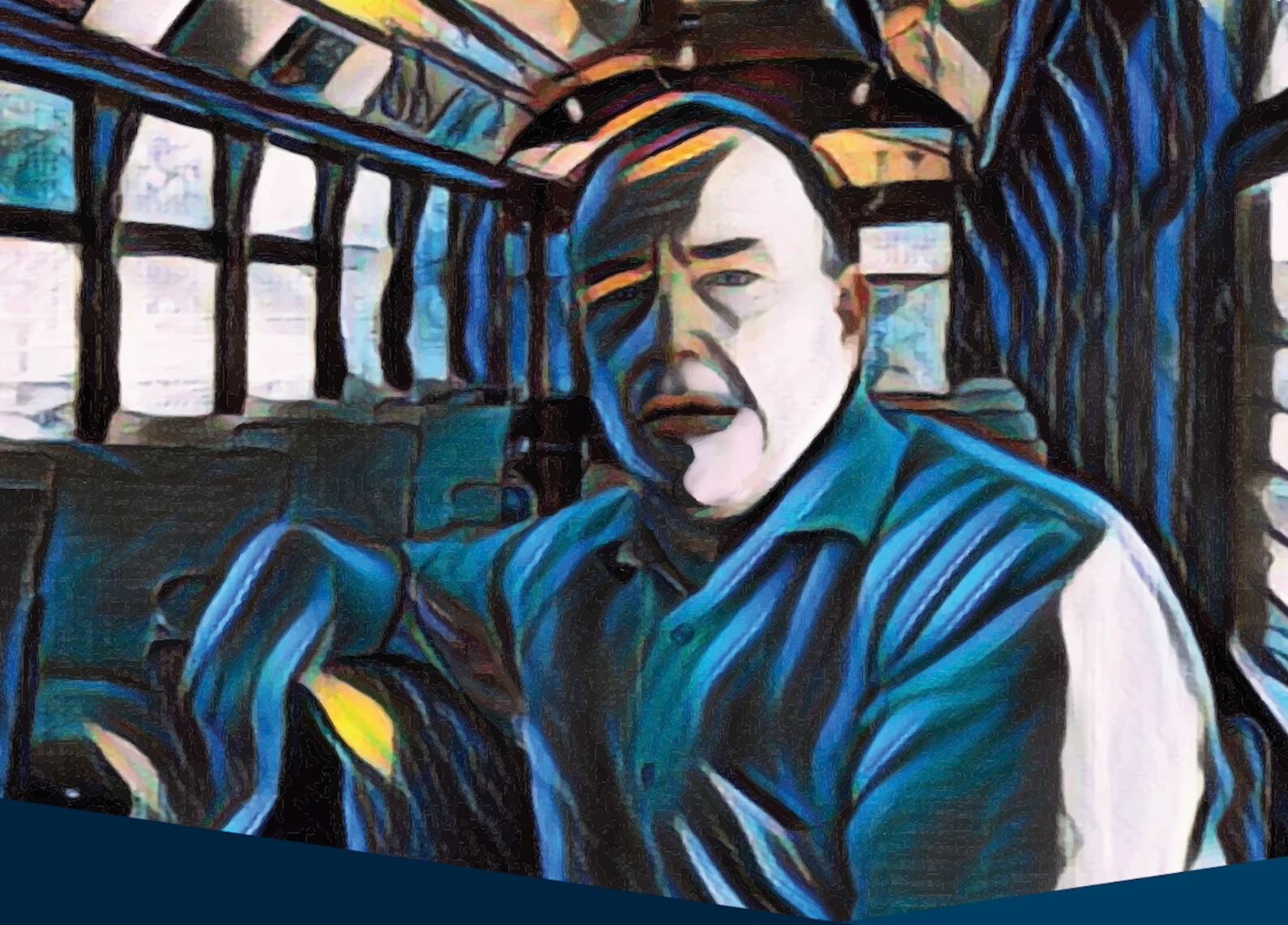


Morfologi, målstrev og maskinar:

Trond Trosterud {fyller | täyttää | deavdá | turns} 60!

Redigert av:
Lene Antonsen
Sjur Nørstebø Moshagen
Øystein A. Vangsnes

NORDLYD
No. 46.1 – 2022
ISSN: ISSN 1503-8599

NORDLYD
No. 46.1

Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!

With abstracts in English

UiT Noregs arktiske universitet/The Arctic University of Norway

TROMSØ 2022

Forsidedesign: Bjørn Hatteng, Grafiske tjenester, UiT Norges arktiske universitet

Forord

Denne utgåva av Nordlyd er eit festskrift til ære for vår kollega professor Trond Trosterud, i samband med at han fyller 60 år 30. august 2022. Utgåva inneheld 22 artiklar skrivne av i alt 43 forfattarar – for det meste folk som har samarbeidd med Trond i tidlegare år. Vi har òg skrive ei innleiing om Trond, og til sist i boka er det ei liste over Trond sine publikasjonar. Vi er umåteleg glade for å kunna heidra Trond med denne boka som inneheld mykje spennande om mange ulike tema frå forskjellige språk og som med det speglar dei vide og mangslungne interessene til jubilanten. Vi takkar alle som har skrive artiklane, og vi takkar alle fagfellane som har vurdert og kommentert artiklane.

Esipuhe

Tämä Nordlydin painos on juhlajulkaisu kollegamme, professori Trond Trosterudin kunniaksi hänen 60-vuotispäivänään 30. elokuuta 2022. Juhlajulkaisu sisältää 22 artikkelia, mitkä on kirjoittanut yhteensä 43 henkilöä - enimmäkseen Trondin kanssa yhteistyötä tehneitä ihmisiä. Olemme myös kirjoittaneet Trondista esittelyn, ja kirjan lopussa on luettelo Trondin julkaisuista. Olemme äärimmäisen iloisia voidessamme kunnioittaa Trondia tällä kirjalla, joka sisältää paljon mielenkiintoista tietoa monista aiheista eri kielistä ja joka tältä osin heijastaa syntymäpäiväsankarin laajaa ja monipuolista perehtyneisyyttä. Kiitämme kaikkien artikkelien kirjoittajia sekä kaikkia kollegoja jotka arvioivat ja kommentoivat näitä artikkeleita.

Ovdasánit

Dát Nordlyd-girji lea ávvočála min bargoustiba professor Trond Trosteruda gudnin, go son deavdá 60 jagi borgemánu 30. b. 2022. Girjjis leat 22 artihkkala maid oktiibuot 43 čálli leat čállán – eatnasat leat olbmot geat leat ovttasbargan Trondain. Girjjis lea maiddái čállosa Tronda birra, ja girjji loahpas lea su almmuhanlistu. Mii leat hirbmat ilus go beassat gudnejahttit Tronda dáinna girjjiin, mas leat ollu gelddolaš čállosat máŋgga fáttás iešguđet gielain, ja dainna lágiin speadjalastá su viiddes ja máŋggasuorggat beroštumiid. Giitit buohkaid geat leat čállán, ja giitit maiddái buot fágaguimmiid geat leat árvvoštallan ja kommenteren artihkkaliid.

Preface

This edition of Nordlyd is a celebratory publication in honor of our colleague, Professor Trond Trosterud, in connection with his turning 60 on the 30th of August 2022. The edition contains 22 articles written by a total of 43 authors - mostly people who have collaborated with Trond in previous years. We have also written an introduction about Trond, and at the end of the book there is a list of Trond's publications. We are immensely happy to honor Trond with this book, which contains a lot of exciting information on many topics from various languages and which in that respect mirrors the wide and manifold interests of the jubilarian. We thank everyone who has written the articles, and we thank all the colleagues who have assessed and commented on the articles.

{Tromsø | Tromssa | Romsa | Tromsø}, {august | elukuu | borgemánnu | August} 2022

Lene Antonsen, Sjur Nørstebø Moshagen, Øystein Vangsnes

Innhald / Content

Innhald ordna etter tema / Content ordered according to topic

Samisk / Saami languages

Kvensk / Kven language

Andre språk / Other languages

Språkplanlegging / Language planning

Språkteknologi / Language technology

Trond Trosterud ved 60

Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes

Trond Trosterud fyller 60 år 30. august 2022. Han er professor i samisk språkteknologi og leiar av forskingsgruppa Giellatekno ved Institutt for språk og kultur, UiT Noregs arktiske universitet. Vegen fram dit har ikkje gått i bein linje, og samisk språkteknologi er på ingen måte det einaste Trond har interessert seg for eller gitt viktige og vektige bidrag til i dei drygt 30 åra han har vore ein aktiv språkforskar. Vi har å gjera med ein breitt orientert og djupt engasjert akademikar med eit stort og mangslunge nettverk både nasjonalt og internasjonalt, noko som tydeleg kjem til uttrykk når ein ser på den tematiske spennvidda i bidraga til dette festskriftet. Også den språklege spennvidda i skriftet er vid. Langt frå alle språk som Trond har kunnskap om og aktive ferdigheiter i, er representerte, men her har vi artiklar på engelsk, bokmål, nynorsk, dansk, finsk, nordsamisk og enaresamisk, og språka som vert undersøkte er langt, langt fleire.

I denne innleiinga vil vi rota litt i fortida til jubilanten – med vekt på det faglege – og også innimellom fortelja litt om når og korleis vi redaktørane vart kjende med han.

Trond Trosterud vart fødd 30. august 1962. Sjølv om han budde dei første leveåra i Oslo og Hammerfest, er det barne- og ungdomsåra på Hundhammeren i Malvik utanfor Trondheim som har sett sitt stempel på talemålet hans: Trond er umiskjenneleg trønder i målet.

Musikk spelte ei viktig rolle i oppveksten til Trond. Han var med i Nidarosdomens guttekor, og Sjur sitt første minne av Trond var ein av dei eldste gutane lengst bak i guttekoret, med stort, krøllete hår, og han vart av mange rett og slett kalla for «han med håret».

Tida på vidaregåande vart formgjevande for resten av livet hans. Då vart han kommunist, og han såg at alkohol for mange ungdomar var ein negativ faktor, så han slutta difor med alkohol då han var 18 år gamal. Han vart også ihuga målmann. Alt dette har han halde på sidan, og livslange venskapsband vart knytte i den tida. Samtidig vart det ideologiske grunnlaget for alt målarbeid, både fagleg og meir privat, lagt i form av interessa for dei svake og utsette og for retten til eige språk. Kampen mot språkleg urett har gått som ein raud tråd gjennom livet hans, og alt starta den gongen.

Men først førte gitarspeling og song han til musikkfolkehøgskolen Toneheim utanfor Hamar og eitt år med musikkstudium i Tromsø. Matematikk og historie studerte Trond også i unge år, men det var språkvitskapen som etter kvart fanga han. Med nordiskfaget i sekken dro han i januar til Åbo for å undervisa eit norskkurs i nokre veker før han skulle avtena verneplikta på Værnes flystasjon, og norskkurset ga meir enn faglege resultat. Ein av studentane var finskspråklege Kirsti frå Kärkölä utanfor Lahtis, og allereie same sommar fekk ho seg sommarjobb i Trondheim så ho kunne vera nær menig Trosterud. I august 1987 flytta dei saman i Oslo og sommaren deretter gifta dei seg. I 1993 vart sonen Sindre fødd og året etter dottera Aino. Finsk vart heimespråket i familien som etter kvart fekk Tromsø som heimebase, og Trond og Kirsti gjekk føre med å krevja finskopplæring for borna sine så dei fekk utvikla både norsk og finsk på ein god måte. I dag er begge borna busette i Finland.

Flytten til Oslo i 1987 representerte starten på den allmennlingvistiske løpebana til Trond. Ved Universitetet i Oslo tok han grunnfag og mellomfag, og i starten var det generativ syntaks og norske dialektar som stod i fokus. Den første artikkelen til Trond frå 1989, om hallingmålet som eit mogleg moteksempel til den såkalla nullsubjektsparameteren (Null Subject Parameter), fekk mykje merksemd i feltet og vart hyppig sitert. Så oppstod ei interesse for bindingsteori og refleksivar i skandinaviske språk og dialektar – med mellom anna ein sampublikasjon med Sjur frå 1990 – og samtidig hadde også kunnskapane

 Nordlyd 46.1: 1–7, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigert av Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publisert ved UiT Noregs arktiske universitet.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6663

i og om finsk auka. Det førte Trond inn i eit hovudfagsprosjekt i lingvistikk om bindingsrelasjonar i to finnmarkfinske dialektar som han gjennomførte ved Universitetet i Trondheim. Og medan det universitetet no er gått opp i den høgare eininga NTNU, ville Trond mykje truleg i dag ha omtalt dei to dialektane frå høvesvis Pyssyjoki (Børselv) og Paatsjoki (Pasvik) som *kvenske* og ikkje finske.

Med den hovudfagsoppåva tok Trond steget over i fennistikken, og etter kvart vart fennistikk til uralistikk. Etter eit par år som norsklektorvikar ved Helsingfors universitet på starten av 1990-talet, fekk Trond ei stipendiatstilling ved Universitetet i Tromsø for å forska på syntaksen i uralske språk. Han hadde vorte del av ei lita gruppe med generative lingvistar frå ulike stader som interesserte seg for finsk syntaks, og ut av samarbeidet deira kom den viktige artikkelen *The structure of INFL and the finite clause in Finnish* frå 1993 med Trond som førsteforfattar (sjå publikasjonslista) saman med mellom andre Anders Holmberg. Det var på denne tida Øystein vart kjend med Trond. Det første møtet var ein kort prat på pauserommet til Institutt for fonetikk og lingvistikk i Bergen – Trond var på veg til eller frå eit forskarkurs på Vatnahalsen – men sommaren 1992 tok Trond imot unge Vangsnes i Helsingfors og viste han rundt: Øystein hadde vore på sitt andre sommarkurs i finsk, og skulle i heile 1993 vera gjestestudent ved Helsingfors universitet. Trond var ein god og raus ven å ha i den finske hovudstaden. Han lét Øystein få låna kontorplassen sin hjå uralistane i den ærverdige hovudbygninga til universitetet, inviterte på tur til svigerfamilien i Kärkölä, diskuterte fag og anna i timesvis, og eit særleg kjært minne er ein lang skitur på Vanda å (Vantaanjoki) frå Forsby (Koskela) og langt innover i Vanda (Vantaa). På den turen viste Trond sin sans for sosiolingvistisk utforsking. Då dei to kakaotørste ville spørja nokon etter ein kafé, sa Trond: «No er vi jo på landsbygda og her er den opphavlege lokalbefolkninga svenskspråkleg, så lat oss sjå om det funkar å spørja på svensk.» Dei så gjorde, og svaret frå den første skiløparen dei møtte, kom på klingande finlandssvensk.

1990-åra kan likevel seiast å vera prega av ei slags akademisk rotløyse for Trond. Midt i tiåret tok han eit par års permisjon for å jobba som urfolkskonsulent ved Barentssekretariatet i Kirkenes. Interessa for uralske språk og fleire turar til Russland for å studera dei hadde gitt han erfaringar og kompetanse som kom vel til nytte der. Men ein stad på vegen hadde Trond mist trua på den generative lingvistikken, og tilbake som stipendiat ved UiT var syntaksforsking vorte til morfologiforsking. Han har sjølv skrive om det estetiske i morfologien, og at noko av det vakraste han veit, er mønstra i finsk grammatikk.

Stipendperioden gjekk ut, og etter eit kort engasjement på det som den gongen var Finsk institutt, baud det seg nye moglegheiter i språkteknologifirmaet LingSoft. Der var Sjur allereie tilsett, og jobben til Trond var å testa grammatikkontrollen deira for norsk bokmål. Utviklingsarbeidet skjedde i nært samarbeid mellom LingSoft og Tekstlaboratoriet ved Universitetet i Oslo. Trond hadde allereie vore med i det fellesnordiske *Samisk datautval* som i 1997 laga tastaturlayout for nord-, enare- og skoltesamisk, og i perioden 1997–2000 var han Noregs representant i ISO/IEC JTC1 WG2, WG3 Character set standardization.

I møtet med språkteknologi såg Trond eit stort potensial for samisk og andre små språk, og med ei kjensle av å falla mellom ulike faglege stolar (nordistikken og fennistikken), bestemte han seg for å gå heilhjerta inn for å jobba for samisk språkteknologi. Den første større løyvinga frå Noregs forskingsråd bar det noko smålåtne namnet «Prosjekt for utarbeiding av samisk språkteknologi» og gjekk føre seg i perioden 2001–2004. I prosjektskildringa stod det at:

> «[Prosjektet] har som hovudmål å utarbeide eit automatisk morfologisk analyseprogram og ein morfologisk disambiguerar for nordsamisk. I tillegg vil prosjektet lage baklengsordbøker, ordformlister, ordformfrekvenslister og maskinlesbare tekstkorpora, som skal publiserast over elektronisk. Innafor ramma av prosjektet vil det og bli utabeidd eit morfologisk analyseprogram for delar av det sørsamiske leksikonet. Desse analyseprogramma vil bli gjort tilgjengeleg for forskarar og andre brukarar over internett.»

Dette var starten på det vi kan kalla Giellatekno-eventyret.

Det var på denne tida Lene møtte Trond for første gong (jamvel om også ho hadde budd to år på Hundhammeren i unge år og dei hadde felles kjende). Ho var på UiT for å følgja undervisning i samisk hovudfagskurs. Dei byrja å drikka kaffi saman i kantina, og etter kvart vart Trond rettleiar for hovudsfagsoppgåva til Lene i samisk språkvitenskap. Via henne kom Trond i kontakt med Samisk språksenter i Manndalen, og dit drog han på språkbad ein sommar for å læra seg munnleg nordsamisk. Medan han var på ferie i Helsingfors, lånte han igjen bort kontorplassen sin, denne gangen til Lene som

fikk det første kjennskapet sitt med språkteknologi då ho i ein sommarjobb la stadnamn frå Statens kartverk si liste inn i analysatoren (ved hjelp av gule postit-lappar med emacs-kommandoar).

Trond arbeidde åleine med samisk språkteknologi fram til 2004 då Divvun-gruppa vart oppretta. I åra etter kom det nokre prosjektstillingar, og Marit Julien og Linda Wiechetek var tidleg med i arbeidet, medan Lene hausten 2006 kom tilbake til UiT for å delta i bygginga av ein syntaktisk analysator for nord- og lulesamisk. I fleire år frå 2002 heldt Trond god kontakt med Tekstlaboratoriet ved UiO, i tillegg til fleire andre liknande miljø i Norden, via PaNoLa-nettverket, *Parsing Nordic Languages*.

Arbeidet med språkteknologi tok det meste av tida og merksemda til Trond i denne tida, men saman med Patrik Bye og Øystein rakk han også å skriva morfologikapittelet i innføringsboka *Språk og språkvitskap* (Samlaget, 2003), og i 2004 gjorde han ferdig dr.art.-avhandlinga si om morfologi i uralske språk: *Homonymy in the Uralic two-argument agreement paradigms*. Begge prøveførelesingane hans er òg gitt ut som vitskaplege artiklar.

Trond i Edmonton, på workshop i regi av ALTLab (Alberta Language Technology Laboratory) ved University of Alberta. Foto: Sjur Nørstebø Moshagen 2015

Frå 2008 vart det oppretta til saman tre faste stillingar i forskingsgruppa Giellatekno. I tillegg til Trond, vart Ciprian-Virgil Gerstenberger og Lene kollegaene hans i faste stillingar, og Linda byrja i stipendiat-stilling same år. Trond var rettleier i alle tre sine doktorgradprosjekt, og Linda og Lene disputerte våren 2018. Trond har også rettleia Robert Reynolds som disputerte i 2016 og Uliana Petrunina i 2018, som begge skreiv om språkteknologi for russisk i avhandlingane sine.

Trond har også undervist mykje, spesielt på masterkurs. Han har rettleidd mange masterstudentar som har skrive oppgåver om og på nordsamisk og sørsamisk, og han har også rettleidd studentar i andre fag, seinast ein lektorstudent på nordisk våren 2022 saman med Øystein.

I arbeidet med språkteknologi var sør-, lule- og nordsamisk i fokus i mange år, men sidan 2014, då Marja-Liisa Olthuis byrja i postdok-stilling ved Giellatekno, har han vore sentral i å bygga opp ein morfologisk analysator for enaresamisk. Dette følgde han opp ved å nytta analysatoren som grunnlag for ordretteprogram, grammatikkretteprogram og maskinomsetjing frå nordsamisk til enaresamisk. Trond har

også vore professor II ved Umeå universitet, og han underviser og rettleier kollegaer og studentar ved Uleåborgs universitet nærmast kvar veke. Sommaren og hausten 2021 var han sentral i bygginga av ein kildinsamisk analysator med grunnlag i Elisabeth Schellers ordboksmateriale og dokumentasjon av grammatikken.

Engasjementet for små språk stoppar ikkje ved samisk. Trond var med i styret for kvensk institutt åra frå 2007 til 2015. Der nytta han posisjonen sin til å saman med instituttet å bygga opp ein analysator for kvensk, og han er framleis mentor og pådrivar i dette prosjektet. Han stiller både på seminar og ved å halda foredrag, og har dessutan halde kurs for tilsette og det kvenske språktinget i språkteknologiarbeidet. Trond har vidare samarbeidd tett med Per Langgård i *Grønlands Sprognævn* og han har vore sentral i arbeidet med å laga eit grønlandsk ordretteprogram. Han har vore med i prosjekt for arbeid med mange fleire språk også. Til dømes vart han særleg invitert til å hjelpa til med å utarbeida ei felles rettskriving for kornisk, og han var medlem av den internasjonale komiteen for standardisering av kornisk i 2007–2008. I 2013 snakka Trond med Antti Arppe om ein liten automat for Plains Cree som Trond hadde laga. Dette vart starten på eit stort samarbeid med University of Alberta om å laga språkteknologi for urfolksspråk i Kanada, og Trond er partnar i prosjektet *21st Century Tools for Indigenous Languages*.

Rundt 2006 var Trond på eit møte i Torshavn, som del av eit nordisk forskarsamarbeid. På heimreisa var det såpass mykje tåke at flyet vart eitt døgn forseinka. I Torshavn hadde Trond som vanleg kjøpt seg ei bok om språk, denne gongen ein færøysk grammatikk, og ventetida på flyplassen brukte han til å starta arbeidet med ein datamaskinell færøysk språkmodell. Arbeidet med denne blei Tronds søndagshobby. På eit visst tidspunkt trong ein på Færøyane ein stavekontroll med open kjeldekode, og då vart Tronds stavekontroll brått aktuell. Så vellukka og omtykt var dette arbeidet at færingane i 2021 ga Trond ein pris for det: *M.A. Jacobsens heiderspris for anna kulturelt verk*, eller Færøyane sin kulturpris, som han òg vert kalla.

Arbeidet med den færøyske analysatoren er eit døme på korleis Trond er proaktiv. Han ventar ikkje på prosjektskildringar og ekstern finansiering, men set i gang, slik at han allereie på første møte kan visa kva som er mogleg å gjera. Leiinga ved UiT liker å be Trond når det kjem gjester frå andre språksamfunn eller akademiske miljø. Når han kjem til møtet, har han alltid på førehand sett seg grundig inn i lingvistikken, forhistoria og moglegheitene for språka som skal diskuterast, og kanskje har han til og med laga ein demoautomat.

Den færøyske kulturprisen er ikkje den einaste Trond har motteke. I 2010 fekk han formidlingsprisen til Fakultet for humaniora, samfunnsvitskap og lærarutdanning ved UiT, og i 2012 tok han saman med Sjur Nørstebø Moshagen imot den nordiske samiske språkprisen *Gollegiella* for arbeidet som Giellatekno og Divvun-gruppene har gjort for samiske språk.

Trond har hatt ei rekkje verv ut over dei som allereie er nemnde over. Frå 2011 til 2020 var han styremedlem i Språkrådet og han var nestleiar i perioden 2016–2019, og han var medlem av styret for Språkbanken frå 2011 til 2017. I perioden 2013–2015 var han president for *The North European Association of Language Technology* (NEALT).

Trond er aktiv på mange felt også på fritida, og vi vil særleg trekka fram arbeidet hans med Wikipedia. Viss ein er interessert i kunnskap om samiske språk og forskingshistoria, så skal ein lesa nynorsk wikipedia, der Trond er hovudforfattar for desse artiklane. Trond har også i periodar vore med i *Wikimedia Noregs* styre, først i 2008, og deretter i 2012.

Vi som er så heldige å vera Tronds kollegaer på same institutt og korridor, kan leggja til dette: Han er optimist, han ser potensialet i oss alle, studentar og tilsette, uansett alder og bakgrunn, han ser at kvar enkelt har ein unik kunnskap fordi dei har den bakgrunnen og dei erfaringane som dei har. Og vi veit at han stiller opp for oss når vi treng det.

Trond er raus og gir av seg sjølv. Viss nokon spør om noko, stoppar han opp og gir gjerne ein time eller to av tida si for å bistå med kunnskapen sin (medan instituttleiing og andre kan vera småfrustrerte over at han ikkje svarer i ein einaste kanal). Han har vore uformell rettleiar for mange, ved at han gjerne diskuterer problemstillingar og korleis ein kan komma seg vidare etter å ha gått seg fast, og det kan skje både på kafferommet og i symjebassenget. Når nokon tar kontakt om ei sak, så kan dei rekna med ein lang e-post som svar, med analyse av problemet og gode råd. Dette har meir enn ein gang ført til at han har hjelpt andre til å skriva så gode søknader om forskingsmiddel at han kanskje sjølv har tapt i konkurransen.

På ei eldre heimeside på UiT har Trond skrive om seg sjølv at han fagleg har gått «i ein stor boge aust, nord og vest, frå norske dialektar via finsk og estisk til andre sida av Uralfjella, og deretter nord og vestover att til samisk». Vi tenkjer at Trond aldri gjekk seg vill, men hadde ein sterk akademisk utferdstrong, og han har med det gjort norsk språkvitskap rikare. Det er mange som kan vera glade for den vandringa han har gjort, og ikkje minst kan det samiske språksamfunnet vera glad for at han enda opp der han gjorde og no i meir enn to tiår har jobba i teneste for dei samiske språka. Og at verktøya som er utvikla av det samiske språkteknologimiljøet ved UiT vert sette stor pris på, er det liten tvil om når ein ser på kor mange som dagleg er inne på tenarane. Vi let likevel innleiinga vår bli avrunda av dette meir lyriske uttrykket for takksemd over dei digitale språktenestene frå Giellatekno og Divvun:

Mun ráhkistan Giellatekno,
Divvun-prográmma,
Divvun-speller-demo ja
Giellatekno Apertium nuvttá
jorgalanprográmma
Giellatekno lea fiidnámus mii lea
Giellatekno lea oavdu

Jeg elsker Giellatekno,
Divvun-programmet,
Divvun-speller-demo og
Giellatekno Apertium gratis
oversettingsprogram
Giellatekno er det fineste som finnes
Giellatekno er et underverk

frå Siri Broch Johansen
Reivvet kommišuvdnii / Brev til kommisjonen.
Utgitt på ČálliidLágádus / ForfatternesForlag.

Vi ønsker deg hjarteleg til lukke med dagen, Trond! / Mii váimmolaččat sávvat dutnje lihku beaivái!

Trond på ein gamal trikk i Edmonton.
Foto: Sjur Nørstebø Moshagen 2014

Mo, *do*, *so*, *da* – duortnussámi dovdomearkan?

Lene Antonsen
UiT Norgga árktalaš universitehta

Abstract

In this article, I examine the dialect forms of a set of North Saami pronouns – *mo*, *do*, *so*, *da* ('I, you, he/she, it'; standardized forms: *mon, don, son, dan*). More specifically, I investigate where the forms are in use and how the forms have developed. The material shows that the final *-n* has changed in a number of stages before it disappeared completely. I suggest that these pronominal forms are a dialect mark of the Torne Saami dialect group (named after the Torne river valley on the border between Sweden and Finland). The pronominal forms are used throughout this dialect area, and the use continues north to Kvænangen in Norway, which in turn belongs to the Sea Sami dialect group. In the Kvænangen dialect there are also a couple of other characteristics that are typical for some of the Torne Saami dialects.

Keywords: North Saami, Torne Saami, dialects, language change

1 Álggahus

Sullii 15 jagi dás ovdal guorahallen mo Gáivuona suopmana morfologiija lei rievdan njealji buolvva áiggis. Trond Trosterud lei dalle mu bagadalli. Ohppen ollu sus dalle go lei mu bagadalli, ja lean oahppan vel eanet sus dan maŋŋel gitta otná rádjai, go letne bargan seammá dutkanjoavkkus. Dán su gudnečoakkáldahkii geavahan liibba loktet áššii masa mun dušše guoskkastin váldofágadutkosis, ja áiggun guorahallat áššii stuorit geahččanguovllus.

Gáivuona suopman gullá duortnussámisuopmana davimus suorgái, muhto suopmanis leat unnán ieš vuođat mat adnojuvvojit duortnussámi dovdomearkan. Mu hypotesa lea ahte pronomenhámit *mo*, *do*, *so* ja *da* leat duortnussámi dovdomearkkat, ja hámit čájehit Gáivuona suopmana gullevašvuođa duortnussámegillii. Gáivuona boares hubmit geavahit daid pronomenhámiid, omd. *Mo lean da viesus* go čállingielas livčči *Mon lean dan viesus*. Dán artihkkalis iskkan gos hámit geavahuvvojit ja maiddái mo hámiid gárggiideapmi boahtá oidnosii materiálas, muhto in čiekŋut gárggideami vejolaš sivaide.[1]

2 Davvisámi suopmanjoavkkut

Sammallahti juohká davvisámegiela golbman suopmanjoavkun: mearrasámegiella, finnmárkkusámegiella ja duortnussámegiella (Sammallahti 1998: 9). Muhtimat gohčodit finnmárkkusámegiela siseatnansámegiellan (omd. Aikio ja Ylikoski 2022) danne go suopman hubmojuvvo maiddái Finnmárkku olggobealde, ja Finnmárkkus hubmojuvvo maiddái mearrasámegiella. Mun guorrasan dán jurdaga ja gohčodan suopmanjoavkku siseatnansámegiellan. Siseatnansámegielas leat guokte stuorra suopmanjoavkku, nuorttabeali ja oarjjabeali suopmanat, main leat stuorra fonologalaš erohusat. Aikio ja Ylikoski gieđahallaba dán guokte variántta guoktin sierra suopmanjoavkun, nu ahte davvisámegielas leat sudno mielde njeallje suopmanjoavkku (Aikio ja Ylikoski 2022).

Mearrasámegiela deháleamos suopmanmearka lea go konsonántaguovddážis leat bisuhuvvon dološ nasála geminähtat, mat muđui davvisámegielas leat šaddan konsonántačoahkkin klusiila + nasála. Mearrasámegielas daddjojuvvo *vuonna* ja *biemmo* dan sajis go siseatnangiela *vuotna* ja *biebmu*, ja suopman

[1] Giittán namahis fágaguoibmeárvvoštalliid go leaba buktán ávkkálaš kommentáraid artihkkalii.

 Nordlyd 46.1: 9–17, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, maid Lene Antonsen, Sjur Nørstebø Moshagen ja Øystein A. Vangsnes leat doaimmahan. UiT Norgga árktalaš universitehta lea almmuheaddji. http://septentrio.uit.no/index.php/nordlyd
https://doi.org/10.7557/12.6394

hubmojuvvo Giehkirnjárggas Ruošša bealde gitta Návunnii[2] earret Porsáŋgguvuonas (Sammallahti 1998: 9–11). Návuonna lea Romssa fylkka davimus vuotna.

Duortnussámegiela suopmanjoavku hubmojuvvo Ruotas, ja de Suomas Eanodaga oarjjimus oasis, ja maid Norggas Ufuohtá rájes gitta Ivgui ja Ráisii (Sammallahti 1998: 9). Sammallahti juohká duortnussámegiela njealljin váldosuopmanin. Lulimusas lea Jielleváris váldosuopman mii hubmojuvvo dušše Ruoŧa bealde. Čohkkirasa suopman hubmojuvvo maiddái Norgga bealde, nugo Ufuohtás, Lulli-Romssas ja Sáčč̄ás. Nubbin davimus suopman leat Gárasavvona suopman mii hubmojuvvo Geaggánvuomis ja Lávnnjitvuomis Ruoŧa bealde ja Ivgus ja Báhccavuonas Norgga bealde. Davimus suopman lea Suomanjárgga suopman mii Sammallahti mielde hubmojuvvo Oarje-Eanodagas Suoma bealde ja Norgga bealde Ivgueanu ja Gálggojávrri rájes gitta Ráisii (Sammallahti 1998: 10, 17–20). Suomanjárgga Norgga beale suopmanat leat unnán dutkojuvvon earret Gáivuona suopman man birra leat čállojuvvon guokte masterbarggu, gč. Eira (2003) ja Antonsen (2007).

Árbevirolaš suopmančilgehus (nugo Sammallahti 1998 ja Jernsletten 2000) ii váldde vuhtii riikkarájiid mat juhket davvisámi guovllu golmma riikii. Riikkaid eanetlogugielat váikkuhit davvisámegillii, ja jus váldá dan vuhtii, de sihke nuortta- ja oarjjábeale suopmanjoavkkuin leat Suoma ja Norgga beliid variánttat ja duortnussámegielas leat Suoma, Norgga ja Ruoŧa beliid variánttat. Dušše mearrasámegiella hubmojuvvo dušše ovtta riikkas, Norggas (Aikio ja earát 2015).

3 Giella rievdá, ja davvisámi sánit leat otnon

Gielat ja suopmanat rivdet dađistaga. Olggobeale áššit sáhttet váikkuhit, nugo ahte eallinvuohki ja kultuvra rievdá, dahje guvlui bohtet olbmot geain lea eará suopman. Jus nuorra buolva ii beasa vásihit váhnemiid ja áhku ja ádjá giela doarvái, dat maid sáhttá dahkat ahte nuoraid giella lea earálágan go váhnemiid giella. Dihto variánttaid árvofápmu sáhttá dahkat ahte dat vuitet eará variánttaid badjel. Muhto gullá maiddái giela lundui ahte rievdá, go jietnadagat váikkuhit nubbi nubbái, ja jietnadagat gahčč̄et oktii dahje jávket. Maiddái dákkár rievdamat levvet olbmuid oktavuođaid bokte.

Sámegielain, nugo máŋggain eará gielain, dáhpáhuvvet fonologalaš rievdamat deattohis stávvalis sáni loahpas. Sáni loahppajietnadagat gahčč̄et oktii, dahje sániid loahppavokála dahje loahppastávval jávká. Maiddái loahppakonsonánta sáhttá jávkat. Diekkár jávkan gohčoduvvo apokopen ja gullá giela lundui. Dien láhkai davvisámegiela sániid loahppa lea rievdan ja otnon ollu áiggi mielde, ja nu lea sáni deaddostávval oktan konsonántaguovddážiin, šaddan dehálažžan go galgá earuhit sániid morfosyntávssalaš mearkkašumi. Lullisámegiella lea bisuhan akkusatiivva *m* ja genitiivva *n* gehčosiid: *jaevrie*, *jaevriem*, *jaevrien* (nominatiiva, akkusatiiva, genitiiva), ja go ii leat dássemolsašupmi de loahppakonsonánta earuha dáid kásushámiid. Hámit leat muhtun muddui otnon julevsámegielas: *jávrre*, *jávrev*, *jávre*, ja davvisámegielas leat loahppakonsonánttat *-m* ja *-n* ollásit jávkan ja dán áigge dušše bárrastávvalmáddagiid konsonántaguovddáš earuha dán guokte kásusa: *jávri*, *jávrri* (nominatiiva, akkusatiiva-genitiiva).

Máŋgga morfosyntávssalaš hámis lullisámegiella lea bisuhan olles stávvalgehčosa mii lea otnon ja neutraliserejuvvon davvi- ja julevsámegielas. Ovdamearkka dihte lullisámi vearbahápmi *båateme*, lea julevsámegielas *boahtám*. Eanaš davvisámegiela suopmaniin lea *-n* ja *-m* gahčč̄an oktii sáni loahpas, ja hápmi lea čállingielas dál *boahtán*.

4 *mo*, *do*, *so* ja *da*-hámiid geográfalaš guovlu

4.1 Maid LIA Sápmi hállankorpus čájeha

Dán artihkkalis guorahalan *mo*, *do*, *so* ja *da*-hámiid davvisámi suopmaniin. Materiálan geavahan vuosttažettiin LIA Sápmi Sámegiela hállangiellakorpusa[3] mas leat jietnabáttit mat leat transkriberejuvvon dála čállingillii. Máŋga suopmanmearkka eai oidno transkripšuvnnas, muhto geavahanlavttas lea vejolaš gulda-

[2] Návuonna-nama čállinhápmi čájeha guovllu mearrasámi suopmana ja lea dohkkehuvvon almmolaš báikenamman (https://stadnamn.kartverket.no/fakta/808281/).

[3] https://tekstlab.uio.no/glossa2/saami (10.11.2021)

lit jietnabáttiid. Materiála boahtá Joho Niillasa (Nils Jernslettena) čoakkáldagas, ja lea báddejuvvon jagiid 1960–1987. Čoakkáldaga 94 hubmi leat buohkat davvisámegielagat, ja sii bohtet 19 báikkis, miehtá Sámi muhto eanaš Norgga bealde. Nuoramus hubmi lei 29 jahkásaš, muhto eatnasat ledje boarráseappot go 60 jagi dalle go jearahallojuvvojedje. Materiálas leat oktiibuot 189.000 sáni. Govvosis 1 leat eanaš báikkit maidda artihkal čujuha.

Govus 1: Kárttas leat báikenamat maidda artihkal čujuha. (Google maps, heivehuvvon)

Materiálas leat eanemustá 1. persovnna pronomenat go hubmi dábálaččat muitala iežas birra. Muhto 1., 2. ja 3. persovnnaid hámit čuvvot ovtta minstara, hupmangielas, *mo*, *do* ja *so,* ja čállingielas *mon*[4]*, don* ja *son*, ja árvvoštalan dan dihte ahte *mo/mon*-pronomen ovddasta buot golbma persovnna. Guorahalan maiddái čujuheaddjipronomena *da* (<*dan*>). Davvisámegiela čujuheaddjipronomenat *dat, dát, diet, duot, dot* čuvvot ovtta sojahanminstara. Čállingielas čujuhit dát pronomenat dábálaččat elliide, dávviriidda ja abstrákta objeavttaide, muhto hupmangielas čujuheaddjipronomen čujuha maiddái olbmuide. Čujuheaddjipronomen sáhttá maid geavahuvvot attribuhttan substantiivii, ja eanaš kásusiin lea dalle olles kongrueansa, nugo *dát girji*, *dán girjji*. Ovttaidlogu illatiivvas ja lokatiivvas lea beallekongrueansa: *dán girjái* ja *dán girjjis*.

Hállankorpusis eai leat juohke guovllus nu galle hubmi, ja báddejuvvon materiála sáhttá leat oalle oanehaš, ja hubmis sáhttet leat variánttat mat eai leat mielde materiálas. Eai leat galle hubmi Ruota bealde, ja ii oktage Jiellevári guovllus mii lea lulimus duortnussámi guovlu. Nubbin lulimus váldosuopman lea Čohkkirasa suopman, masa gullá Skániid suopman. LIA materiálas leat dán guovllus golbma hubmi, ja sin gielas leat variánttat. Dábálaččat ii leat loahppa *-n* go čuovvovaš sátni álgá konsonántan, nugo *mo val* (1a) ja *da guovllun* (1c), ja lea loahppa *-n* go čuovvovaš sátni álgá vokálain, *dán áiggin* (1b).

(1) Hubmit riegádan áigodagas 1910–1920, Skániin (LIA)

a. in **mo** val ipmir olugiid [...]

b. dah leat muhtimat **dán** áiggin geah [...]

c. **da** guovllun orru

Sáččās lea okta hubmi, riegádan 1905:s. Su suopmanis leat pronomenat *mo* ja *da*. Loabágis lea okta hubmi, gii lea badjeolmmoš, ja su hupmamis eai leat duortnussámi suopmanmearkkat. Sus lea čielga siseatnan-

[4] LIA-transkripšuvnnain lea dábálaččat čállojuvvon *mun* vaikko hubmi dadjá *mo* dahje *mon.*

suopman ja son geavaha persovdnapronomena *mun*. Ruoŧa bealde leat dušše guokte hubmi geat gullet Čohkkirasa váldosuopmanii, Gironis ja Čohkkirasas, ja sudnos leat *mo* ja *da*-hámit.

Nubbin davimus duortnussámi váldosuopman lea Gárasavvona suopman, mii maiddái hubmojuvvo Norgga bealde. LIA materiálas dán váldosuopmanis leat dušše guokte hubmi. Nubbi lea Gárasavvonis ja sus gielas lea *mo*, *do*, *da* (<*mon*>, <*don*>, <*dan*>). Nubbi hubmi lea badjeolmmoš Rátnáhis geas árvideames lea gullevašvuohta maiddái Gáresavvona guvlui. Son dadjá *dan* ovddabeale vokála (2a), ja *da* ovddabeale konsonántta (2b), muhto persovdnapronomena 1. persovnna nominatiivahápmi lea dušše *mo* (2a) beroškeahttá čuovvovaš sániid álgojietnadagas.

(2) Hubmi Rátnáhis (LIA)

a. **dan** in dieđe gal **mo**

b. **da** rájin gal heaitán johtin

Davimus duortnussámi váldosuopman lea Suomanjárgga suopman. LIA-materiálas leat guokte hubmi Omasvuonas, mii lea Suomanjárgga ja Gárasavvona váldosuopmaniid ráji alde. Sudno hupmamis gullo *mo* ja *da*. Gáivuonas leat guhtta hubmi, ja maiddái sis gullo *mo* ja *da*. Ovtta hupmis gullo *da* ovddabeale konsonántta: *da stuora* (3a) ja *dan* ovddabeal vokála: *dan áigge* (3b), muhto persovdnapronomeniid nominatiivahámit leat loahppa *-n* haga maiddái ovddabeal konsonántta: *do it leat* (3c).

(3) Hubmi, riegádan birrasiid 1920, Gáivuonas (LIA)

a. **da** stuora ábi nalde

b. lohpi bivdit **dan** áigge go

c. go **do** it leat oahppan bar- gáddebarggu bargat

Davimus guovllus gos dát suopmanmearka gullo LIA-materiálas, lea Návuonas. Návuonas leat golbma hubmi[5], riegádan áigodagas 1901–1907. Buot golmma návuotnalačča gielas leat mearrasámi dovdomearkkat go leat bisuhuvvon dološ nasála geminähtat, nugo *eanni* (<*eadni*>), ja palatála klusiila sajis gullo palatála friktatiiva: *dajjá* (<*dadjá*>). Hubmit geavahit 1. persovnna nominatiivan sihke *mo* (4a, 4c) ja *mun* (4b), muhtun muddui beroškeahttá čuovvovaš sáni álgojietnadagas. Čujuheaddji pronomen genitiivahápmi orru dávjjit čuovvumin minstara *dan* ovddabeale vokála (4d), ja *da* ovddabeale konsonántta (4e).

(4) Golbma hubmi, riegádan birrasiid 1910, Návuonas (LIA)

a. vaikko **mo** oačču̇n práhtet sámegiela

b. **mun** leamaš báktebarggus

c. nei **i mo** gal muitte dál

d. **dan** áiggi

e. dat **da** rájes go lean álgán bargat gitta

LIA-korpusa eará guovlluid hubmit dadjet *mun/mon* ja *dan*.

4.2 Eará materiálat

LIA-korpusis ii leat oktage hubmi lulimus davvisámi guovllus. Grundström pronomentabeallas (1946–1954: 1745) leat davimus hubmit Girjásis eret, ja sin hupmangielas leat hámit *mȯn, tȯn* ja *sȯn* ja ovddabeale konsonántta *mȯ*, *tȯ* ja *sȯ*.

Čohkkirasa suopmana leat moattis dutkan. Eriksen (2009) čállá Norgga beale suopmana birra ahte ovttaidlogu nominatiivvas dávjá gullo *mo*, *do* ja *so*. Sihke Collindera (1949: 239–240) ja Jernslettena (2000: 60) mielde leat ovttaidlogu nominatiivahámit main lea, ja main ii leat loahppa *-n*. Collinder (1949: 250) čállá ahte čujuheaddji pronomena akkusatiivva-genitiiva hápmin geavahuvvo *dan* (*tan ~ ðan*) ovddalbeal vokála ja *da* (*ta ~ ða*) ovddabeal konsonántta.

[5] Návuonas lea vel okta hubmi, muhto sus lea báddejuvvon dušše moadde celkosa.

LIA-materiálas eai leat hubmit Ivgu suohkanis. Nesheim (1962: 347) namuha ahte sihke Moskavuonas ja Ivgus nominatiiva ovttaidlogu persovdnapronomeniin leat variánttat, main lea ja ii leat loahppa *-n*.

5 Hámiid gárggiideapmi

LIA-materiála ja čálalaš gáldut maidda lean čujuhan mannan kapihttalis, čájehit ahte *mo*, *do*, *so* ja *da* hámit leat olles duortnussámi guovllus, ja dasa lassin Návuona suopmanis. Dán kapihttalis áiggun geahčča lagabut mo hámit *da* ja *mo* leat gárggiidan. Čájehan dihte giellarievdamiid, de geavahan maiddái oasi materiálain maid čoggen iežan váldofágabargui mas guorahallen dihto morfologalaš rievdamiid Gáivuona suopmanis. Materiálas leat máidnasat maid guhtta diehtoaddi riegádan áigodagas 1855–1877 leat muitalan Qvigstadii (QLES), ja su transkriberemiin sáhttá muhtun muddui oaidnit sin suopmana. Materiálas lea dasa lassin okta informánta riegádan 1870-logus, son orui Moskavuonas. 14 informántta leat riegádan áigodagas 1884–1914, ovcci informántta riegádan 1926–36 ja gávccis riegádan 1955–1972. Nuoramus buolvva informánttain lea sámegiella 2. giellan. Dainna materiálain sáhttá guorrat mo giellarievdamat leat dáhpáhuvvan njealji buolvva áiggis.

Hubmi riegádan 1890, Moskavuonas eret, geavaha sihke *mo* ja *mon*, muhto go guorahallá dárkileappot de boahtá ovdan, nugo ovdamearkacealkagiin (5), ahte loahppa *-n* jávká dušše go maŋit sátni álgá konsonánttain. Čujuheaddji pronomenis dáidá čuovvut seamma minstara, go son dadjá *da gaskka* (5c) ja *dan áiggi* (5d). Su gielas maiddái genitiiva pronomen (<*mu*>) čuovvu seamma minstara, go dadjá *mūn áhčči* (5e) (<*mu áhčči*>) ja *mū gieđaide* (5f).

(5) Erik Eriksen, Moskavuonas eret, riegádan 1890 (Nesheim lea bádden)

a. dah lei vuosttaš mašiidna maid **mon** oidnen

b. [...] jus **mon** in leat siiddan

c. **da** gaskka go **mo** lean dáppil [...] de **mo** bohten siidii [...]

d. mii **dan** áiggi gohčodedje eksirsan

e. **mūn** áhčči dah lea maiddái leamaš čađah áiggi bivdin

f. goappaš **mū** gieđaide

Qvigstada máinnasteaddjit leat boarráseappot, ja sin gielas leat loahppa *-n* dušše muhtumin jávkan ovddabeale konsonántta. Čuovvovaš guokte diehtoaddi jietnadeaba *-n* maid dalle go čuovvovaš sátni álgá konsonánttain (6a), muhto sudno gielas leat maiddái dáhpáhusat main lea n-apokope konsonántta ovddabealde (6b).

(6) Erik Persen, riegádan 1856, Gáivuonas eret

a. Mait **don** siđat ad'det [...] **Mon** siđalin [...]. [...] ige **son** diettan [...] (QLES III: 194)

b. Gosa **do** læk manname? [...] (QLES III: 198)

Gáivuona boarráseamos báddejuvvon diehtoaddiid gaskkas leat hubmit geain leat sihke *mo* ja *mon* ovddabeale vokála, muhto sii eai geavat *mon* ovddabeale konsonántta, muhto *mo* geavahuvvo maiddái ovddabeale vokála (7). Dát čájeha ahte *mo* hápmi vuoitigoahtá.

(7) Hubmi riegádan 1884 Gáivuonas (Nesheim lea bádden)

a. Ja eanet **mon** in muitte dan ... skuvlla birra muitalit.

b. Ja Finnamárkkus **mon** in muitte man olu **mo** doppe lean johtán.

c. de **mo** uožžun báhpas

Hubmiid giella čájeha nuppástuvvanproseassa. Ođđa hápmi man loahppa *-n* lea jávkan, ihtá árbevirolaš hámi báldii, álggos dušše dalle go čuovvovaš sátni álgá konsonánttain, muhto ođđa hápmi orru vuoitimin eanet ahte eanet. Čuovvovaš buolvva hubmit leat riegádan 1930-logus, ja sin gielas measta ii gávdno

loahppa *-n* persovdnapronomeniid nominatiivahámis. Sis lea *n*-apokope beroškeahttá čuovvovaš sáni álgojietnadagas (8), ja nu lea giellarievdan dáhpáhuvvan ollásit.

(8) Guokte hubmi riegádan 1936 Gáivuonas (Antonsen lea bádden)

a. **mo** lean riegádan dás

b. **mo** in gal dieđe mii dat lei

c. ja gávnnai olggos aht **so** galgá goađi dahkat

d. nu ahte **so** ii dieđe gal gos dat lea

Čujuheaddji pronomen akkusatiiva-genitiiva ovttaidlogu hámis leat boarráseamos materiálas golbma hámi: *dam, dan* ja *da.* Qvigstada máinnasteaddjit muhtun muddui earuhit dán pronomena akkusatiiva- ja genitiivakásusiid, *dam* ja *dan* (9a ja 9b). Dát hámit geavahuvvojit maiddái lullisámegielas, mas dát guokte kásusa velge earuhuvvojit, akkusatiivahápmi *dam*, ja genitiivahápmi *dan*. Maiddái julevsámegielas earuhuvvojit kásusat, *dav* ja *dan*. Davvisámegielas leat loahppa *-m* reduserejuvvon ja šaddan *-n* ja dát guokte kásusa leat gahččan oktii. Siseatnansuopmanis ja čállingielas loahppa *-n* lea bisson. Maiddái eará nomeniid akkusatiiva ja genitiivahámit leat gahččan oktii, go loahppakonsonánta lea oalát jávkan, muhto dat lea dáhpáhuvvan árabut, nugo *dam bađi* gihpus (9), mas čujuheaddji pronomenis lea akkusatiivageažus, muhto substantiivvas ii leat.

Sammallahti oaivvilda ahte substantiivvaid genitiiva loahppanasála jávkan ii dárbbaš leat nu boares dáhpáhus, go davvisámegiela oarjjabeali suopmaniin leat advearbavariánttat *nala ~ ala* ja *nalde ~ alde* maid álgobustávva *n* álgoálggus lei attribuhta genitiivageažus (**várin alde* > *vári nalde*) (1998: 65–66).

Dán artihkkala materiálat čájehit ahte čujuheaddji pronomeniid loahppa *-n* jávká ovtta buolvva maŋŋeleappos go persovdnapronomeniin. Hubmi gii lea riegádan 1912:s, muhtun muddui bisuha čujuheaddji pronomena loahppa *-n* ovddabeale vokála (10), ja hubmiin geat leat riegádan 1950-logus, loahppa *-n* lea oalát jávkan (11).

(9) Erik Persen riegádan 1856:s Gáivuonas.

Vuowde sudnji **dam** bađi, mi ieš duol'da. (QLES III: 358)

(10) Hubmi riegádan 1912:s Gáivuonas (Nesheim lea bádden)

[...] mii diehtit **dan** ahte dohko mii mannat. Muhte **da** mii eain dieđe goassege ahte boahtitgo doppe olggos heakkas

(11) Hubmi riegádan 1950-logus, Gáivuonas (Antonsen lea bádden)

a. Mo šadden bajás **da** áiggis go mii eain galgan oahppat sámigiela.

b. Mo jáhkán aht' golai meid **da** guvlui ahte [...]

Dát giellarievdamat leat tabeallas 1 govviduvvon dainna lágiin: Vuosttaš ceahkis pronomenhámiin lei loahppakonsonánta. Čujuheaddji pronomeniin lei boarráseamos ovdamearkkain maiddái erohus akkusatiivva ja genitiivva gaskkas, ja de loahppakonsonánttat sudde oktii ja boađus lei *-n*. Nuppi ceahkis jávkkai loahppa *-n* go čuovvovaš sátni álgá konsonánttain, ja goalmmát ceahkis lea *-n* jávkan ollásit.

	Vuosttaš ceahkki	**Nubbi ceahkki**		**Goalmmát ceahkki**	**Normerejuvvon čállingiella**
		ovddabeale vokála, ja celkosa loahpas	***ovddabeale konsonántta***		
1	Persovdnapronomeniid nominatiivahámit: *mon don son*	*mon don son*	*mo do so*	*mo do so*	*mun ~ mon* *don son*
2	Čujuheaddji pronomeniid akkusatiiva ja genitiiva: *dam dan > dan*	*dan*	*da*	*da*	*dan*

Tabealla 1: Duortnussámegiela pronomeniid loahppa -n jávkan cehkiid mielde.

Vaikko loahppa *-n* gárggiideapmi lea dáhpáhuvvan seamma láhkai namuhuvvon pronomenhámiin, de dat ii leat dáhpáhuvvan olles gielas oktanaga. Loahppa *-n* jávkkai vuos persovdnapronomeniid nominatiivahámiin, ja veaháš maŋŋelis čujuheaddji pronomeniin. Vejolaččat maŋit giellarievdan lea dáhpáhuvvan analogiijan, go sániid fonologalaš hámit sulastahttet nubbi nuppi: okta stávval ja monoftoŋŋa. Maiddái eará diekkárlágan pronomeniin leat dáin suopmaniin variánttat: *mi ~ min, mū ~ mūn*, gč. (5e, 5f) ja *man* sánis: *man ~ ma*. Maiddái biehttalanvearbba *in*-hámis jávká loahppa *-n* ovddabeal konsonántta: *i mo* (<*in mon*>).

6 Hámit *mo*, *do*, *so* ja *da* duortnussámi suopmanmearkan

Vaikko Gáivuona suopman gullá davimus duortnussámi váldosuopmanii, de Gáivuona suopman sulastahttá maiddái siseatnansuopmana, erenoamážit Guovdageainnu suopmana (Eira 2003). Nesheim (1962) čállá ahte Gáivuona suopman sulastahttá Guovdageainnu suopmana eanet go Ivgu-guovllu sámegiela suopman muđuid dahká. Bárrastávvalmáddagiid lokatiiva ovttaidlogu *n*-geažus adnojuvvo duortnussámegielas dovdomearkan (Sammallahti 1998: 10), muhto lokatiiva kásusa geažus sáhttá Gáivuonas leat sihke *-n* ja *-s*, ja dávjjit *-s* go *-n* (Antonsen 2008), juoga mii láivuda Gáivuona suopmana gullevašvuođa duortnussámegillii. Maiddái LIA korpusa Omasvuona hubmiin gullo dávjjit lokatiiva *-s* go *n*-geažus bárrastávvalmáddagiin. Lokatiiva *-n* gullo erenoamážit báikenamain, mat leat konservatiiva oasit gielas, omd. *Horsnessan*. Báikenamat leat konservatiiva giellaoasit, ja dán sáhttá dulkot nu ahte lokatiiva *n*-geažus lea leamaš dábálaččat áiggis ovdal go hubmit báddejuvvojedje.

Omasvuonas ja Gáivuonas leat muđuige unnán dain suopmanmearkkain maid Sammallahti logahallá duortnussámi dovdomearkan (1998: 17–20). Duortnussámi suopmanmearkkaid gaskkas Sammallahti namuha maiddái deaddostávvala guhkes vokála sekundára diftoŋŋiserema, go gullo *buohten* ja *giessen* (<*bohten*> <*gessen*>), mii su mielde dušše gullá Gárasavvona suopmanii. Dát suopmanmearka lea maiddái merkejuvvon Gáivuona diehtoaddiid gielas QLES-materiálain, ja gullo muhtun hubmiin Gáivuona-materiálas, ja dasa lassin Omasvuona LIA-hubmiin. Eará mihtilmas duortnussámi suopmanmearkkat eai gullo ollenge Omasvuona rájes davás, nugo ahte nuppi stávvala guhkes *i* ja *u* jietnaduvvojit *ie* ja *uo*: *askie*, *viessuo* (<*aski*> <*viessu*>), dahje nuppi stávvala *a* jietnaduvvo *uo* go vuosttaš stávvala vokála lea *o* dahje *u*: *tolluo* ja *ruhtuo* (<*dolla*> <*ruhta*>).

Mu hypotesa lei ahte hámit *mo*, *do*, *so* ja *da* leat mihtilmasat duortnussámi suopmaniidda. Hypotesa doallá deaivása, go hámit geavahuvvojit Jiellevári rájes gitta Gáivutnii. Muhto hámit gullojit maiddái Návuonas materiálas, mas lea mearrasámi buot deháleamos dovdomearka: dološ nasála geminähtaid bisuheapmi. Mearrasámegiela suopmaniin muđui ii gullo *mo*, *do*, *so* ja *da*, go dat goit ii leat Liidnavuona materiálas (Henriksen 2002: 130), iige leat Qvigstada Gállafierdda ja Áillu materiálas (Qvigstad 1925), iige leat LIA-korpusa Unjárgga materiálas.

Návuona suopmanis lea maiddái nubbi dovdomearka mii gullo duortnussámi suopmaniin (ja julev- ja lullisámegielas): sátni *gait*[6], mii lea amas siseatnansuopmaniin. Vearbbaid 1. persovnna máŋggaidlogu preterihttahámi *-jn* geažus, *váccijn* (<*vácciimet*>), gullo LIA-korpusis Omasvuona gielas gitta Návunnii.

[6] **gait** pron. all allt alla, **gait** adv. allra (Svonni 2013: 85).

Maiddái dán vearbahámis lea dáhpáhuvvan apokope, go olles loahppastávval lea jávkan, ja dasto lea loahppa *-m* šaddan *-n*. Dát vearbahápmi gullo maiddái Liidnavuonasuopmanis (Henriksen 2002: 130), mii adnojuvvo mearrasámi suopmanin.

Maiddái Omasvuonas ja Gáivuonas olbmot dáidet dolin hupman mearrasámegiela. Ivgu miššoneara Andreas Sommer raporterii jagi 1771 Roandima bismái[7] ahte olmmošlohku lei lassánan hirbmadit Ivgus, mii dalle govčča maiddái Omasvuona ja Gáivuona. Olbmot ledje muitalan sutnje ahte olmmošlohku lei 60 jagis šaddan 10 geardásažžan. Sommer raporterii ahte ollu geafes badjeolbmot ledje boahtán Ivgui, erenoamážit Ruoŧa bealde (Lindbach 2019: 177). Ivguvuovdi lea lunddolaš johtingeaidnu badjeolbmuide geat bohte Ruoŧa bealde, ja dát sápmelaččat sáhttet áiggi mielde leat jávkkahan mearrasámegiela Ivgu rájes davás. Maiddái Návuonas lassánii olmmošlohku seamma áigodagas, muhto ii seamma olu. Olmmošlohku šattai golmmageardásažžan 1716 rájes gitta 1800 rádjai, ja 90 % ledje sápmelaččat (Bjørklund 1985: 147–148).

Sámi bivttasvierru čájeha mo olbmot leat johtán ja sin oktavuođaid. Duortnussámi nissongahpir mii gohčoduvvo *gobbágahpirin*, *deavddagahpirin* dahje *duorramin*, lea geavahuvvon Čohkkirasa rájes davás. Gahpir lea geavahuvvon maiddái Gáivuonas (Antonsen 1995) ja davás gitta Návunnii, ja Láhpi rájes nuorttas nissonat leat geavahan earálágan gahpira (Fors ja Enoksen 1991: 46–48). Nu nissongahpir čájeha ahte duortnussámit leat johtán gitta Návunnii, ja sii ledje oassin olmmošlassánemis mii dáhpáhuvai 1700-logus. Dát čilgešii manne Návuona suopmanis leat sihke mearrasámi ja duortnussámi suopmanmearkkat.

7 Loahppasánit

Dán artihkkalis lean guorahallan mo sátnehámit *mo*, *do*, *so* ja *da* leat gárggiidan, ja gos hámit geavahuvvojit. Miellagiddevaš lea ahte persovdnapronomeniid loahppa *-n* lea jávkan veaháš ovdal seammasullasaš gárggiideapmi lea čuohcan čujuheaddji pronomeniidda ollásit. Materiálas leat máŋga ovdamearkka das ahte hubmis lea persovdnapronomeniid *n*-apokope, muhto čujuheaddji pronomeniid loahppa *-n* lea jávkan dušše go čuovvovaš sátni álgá konsonánttain.

Suopmanat Omasvuona rájes davás váillahit eanaš daid duortnussámi dovdomearkkaid mat loga-hallojuvvojit earuheaddjin eará suopmaniid ektui. Muhto dain leat dattetge hámit *mo*, *do*, *so* ja *da* maid sáhttit atnit duortnussámi suopmanmearkan.

Referánssat

Aikio, Ante, Laura Arola ja Niina Kunnas. 2015: Variation in North Saami. – Smakman, D. ja Heinrich, P., *Globalising Sociolinguistics: Challenging and Expanding Theory*. Routledge. s. 243–255. Olámuttos: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315697826-32/variation-north-saami-ante-aikio-laura-arola-niina-kunnas.

Aikio, Ante (Luobbal Sámmol Sámmol Ánte) ja Jussi Ylikoski. 2022. North Saami. Girjjis: Marianne Bakró-Nagy, Johanna Laakso ja Elena Skribnik (doaimm.), *The Oxford Guide to the Uralic Languages*. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198767664.003.0010 .

Antonsen, Lene. 1995. *Sjøsamisk klesbruk i gamle Lyngen*. Gáivuona NSR/Kåfjord sameforening.

Antonsen, Lene. 2007. *Giella buolvvas bulvii. Gáivuona sámegiela morfologiija guorahallan*. *Sámegiela váldofágadutkamuš*. UiT Romssa universitehta. Olámuttos: *https://hdl.handle.net/10037/1210* .

Antonsen, Lene. 2008. Lokatiivva gehčosat Gáivuona suopmanis. [English summary: Locative suffixes in the Gáivuotna dialect.] *Sámi dieđalaš áigečála* 2/2008: 3–20. Olámuttos: https://site.uit.no/aigecala/sda-2-2008_antonsen/.

Bjørklund, Ivar. 1985. *Fjordfolket i Kvænangen. Fra samisk samfunn til norsk utkant 1550–1980.* Universitetsforlaget AS, Oslo.

[7] Journal for Aarene 1769 og 1770, insendt in duplo, til Hds Høyædle Høyærværdighet Biskopen over Trondhiems Stift Dr Johann Ernst Gunnerus» Biskopen i Nidaros 81b, Statsarkivet i Trondheim (Lindbach 2019: 177 bokte).

Collinder, Björn. 1949. *The Lappish dialect of Jukkasjärvi. A morphological survey*. Skrifter Kungl. humanistiska vetenskapssamfundet i Uppsala, Vol. 37:3. Almqvist & Wiksell, Uppsala.
Eira, Inger Marie Gaup. 2003. *Giella vákkis vággái. Gáivuona dialeavtta suokkardallan*. Dieđut 2003/2. Sámi instituhtta, Guovdageaidnu.
Eriksen, Ardis Ronte. 2009. *"Mon val vuorrástuvam duoinna nuortasámegielain" Ufuohta ja Oarje-Romssa suopmana gullevašvuohta davvisámegillii ja julevsámegillii*. Sámegiela masterdutkamuš, Romssa Universitehta. Olámuttos: https://hdl.handle.net/10037/1933.
Fors, Gry ja Ragnhild Enoksen. 1991. *Vår folkedrakt. Sjøsamiske klestradisjoner*. Sámi Instituhtta/Davvi Girji o.s., [Kárášjohka/Guovdageaidnu].
Grundström, Harald. 1946–1954. *Lulelapsk ordbok: Lulelappisches Wörterbuch*. Lundequistska bokhandeln, Uppsala.
Henriksen, Marit B. 2002. *Liidnavuona suopmana fonologiija. Mearrasámegiela suopmana fonologalaš guorahallan.* Sámegiela váldofágadutkamuš. Tromssa universiteahtta.
Jernsletten, Nils. 2000. *Davvisámi suopmanat.* Dialektkompendium, samisk grunnfag. Humanistisk fakultet. UiT.
Lindbach, Harald H. 2019. *Minoritetspolitiske dokumentasjonsstrategier i Nordområdet på 1700-tallet. En komparativ analyse av hvordan og hvor samer og kvener trer frem i arkivene til lokal- og regionalforvaltningen i Danmark-Norge og Sverige, med spesielt blikk på Nord-Troms, Jukkasjärvi og Enontekis.* Avhandling levert for graden philosophiae doctor. Olámuttos: https://hdl.handle.net/10037/18927.
Nesheim, Asbjørn. 1962. The Lappish dialect of Ullsfjord and its relation to other Lappish dialects. Girjjis: *Commentationes fenno-ugricae. In honorem Paavo Ravila.* Mémoires de la Société Finno-Ougrienne 125, s. 333–360. Suomalais-ugrilainen seura, Helsinki.
QLES = Qvigstad, Just K. 1929: Lappiske eventyr og sagn III–IV. Instituttet for sammenliknende kulturforskning, Oslo.
Qvigstad, Just K. 1925. *Die lappischen Dialekte in Norwegen: Lappische Texte aus Kalfjord und Helgöy. Reste eines ausgestorbenen Seelappendialektes*. Oslo Etnografiske Museums skrifter, Bind 1 Hefte 1. Oslo.
Sammallahti, Pekka. 1998. *The Saami languages. An Introduction*. Karasjok: Davvi Girji.
Svonni, Mikael. 2013. *Sátnegirji : davvisámegiela-ruotagiela, ruotagiela-davvisámegiela = Ordbok : nordsamisk-svensk, svensk-nordsamisk*. ČálliidLágádus, Kárášjohka.

All that glitters …
Interannotator agreement in natural language processing

Lars Borin
University of Gothenburg

Abstract

Evaluation has emerged as a central concern in natural language processing (NLP) over the last few decades. Evaluation is done against a *gold standard*, a manually linguistically annotated dataset, which is assumed to provide the ground truth against which the accuracy of the NLP system can be assessed automatically. In this article, some methodological questions in connection with the creation of gold standard datasets are discussed, in particular (non-)expectations of linguistic expertise in annotators and the interannotator agreement measure standardly but unreflectedly used as a kind of quality index of NLP gold standards.

Keywords: Evaluation, natural language processing, interannotator agreement, annotation

1. Introduction

Church and Hestness (2019) present "[a] survey of 25 years of evaluation" in speech and language processing.[1] The *evaluation* that they refer to – and this is how this notion is most commonly understood in present-day NLP – involves the following two elements:[2] (1) an NLP system/application for automatically annotating text data for some linguistic features (part of speech, syntactic structure, sentiment polarity, etc.); and (2) a set of text data manually or semi-manually annotated for the same linguistic features, which has not been used in the development of the NLP system to be evaluated. Such a dataset is referred to as a *gold standard* in the literature.

The title of Church and Hestness (2019) refers to a development that the field of NLP has undergone over the last three decades. Trond Trosterud and I both started out in NLP in the previous millennium, and we also have in common an academic background in linguistics rather than in computer science, which has informed the professional trajectories of both of us in the field. At that time, linguistics played a larger role in NLP than at present and rule-based ("symbolic") solutions explicitly realizing linguistic formalizations formed the bulk of NLP systems.

Much has changed since then, and the field looks very different now. Over the last two decades or so, NLP has become increasingly disassociated from the concerns of linguistics (see, e.g. Reiter 2007, Wintner 2009, Manning 2015), and at present, data-driven machine learning approaches hold sway in our field, in particular so-called deep learning (Manning 2015). This development has happened in parallel with the increasing emphasis on evaluation, with noticeable effects on the conception of how gold standard preparation should be carried out.

The same gold standard datasets that are employed for evaluation are also frequently used in order to build NLP systems, in particular systems based on machine learning, such that part of the data is used for training, another part for testing and yet another part for evaluation.[3] Here, we are concerned only with evaluation, however, since this is something that in fact impacts all kinds of NLP systems, regardless of their underlying architecture. In this way the introduction of systematic evaluation of systems is a more profound change in NLP than the "non-symbolic revolution" whereby rule-based applications have been replaced by systems based on machine learning. Thus, the rule-based systems described by Antonsen et al. (2010) and Harrigan et al. (2017) are also formally evaluated against gold standards.

[1] Here I will be concerned only with computational processing of text (not speech processing), a field which goes under several names: *computational linguistics* (CL), *natural language processing* (NLP), *language technology* (LT), and *(natural) language engineering* ([N]LE). In this article, I will refer to it as "NLP". Note that in popular media, much of what is today referred to as "artificial intelligence" (AI) or even "algorithms" is in fact NLP.

[2] In this article, I will discuss only "intrinsic" evaluation, where the output of an NLP system is evaluated directly against an annotated dataset, and not the "extrinsic" kind where such a system is evaluated for its contribution to solving some external, "downstream", task, e.g. a lemmatizer used as a component in an information-retrieval application. For lack of space, I will also need to forgo human evaluation of NLP-system output, which presents plenty of interesting methodological challenges of its own (cf. Hämäläinen and Alnajjar 2021). Further, my interest here is in annotations conforming to best-practice linguistic analysis, rather than to, e.g., popular (mis)conceptions of language. The latter is of course a perfectly valid and interesting research topic, but not the one in focus here.

[3] This presupposes very large datasets, however, and there are many gold standards which can only be used for evaluation because of their size.

©2022 Lars Borin. *Nordlyd* 46.1: 19-26, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6348

Since evaluation is *de rigueur* in today's NLP, the quality of gold standard datasets should be a high priority in the NLP community. Regrettably, much remains to be desired in this regard, and arguably, this is largely due to how the annotation of gold-standard datasets is carried out. This was pointed out already a decade and a half ago by Zaenen (2006), who specifically pointed to the lack of linguistic analysis in NLP annotation practices. Evaluation has taken center stage in NLP since then, but the development of machine learning has partly redefined the role of linguistic analysis in this context, as we will see below. We should also not forget that datasets with even quite basic linguistic annotation are still lacking for the vast majority of the world's some 7,000 languages (Trosterud 2006; 2012), making linguistic annotation a high-priority, long-term concern.

The point about annotation quality raised above was brought home to me serendipitously about ten years ago, when I happened to be present at a presentation by Anthony Kroch, a historical linguist at the University of Pennsylvania, who had been a pioneer in using diachronic treebanks for studying English syntactic change. He had been involved in the compilation of several such treebanks using the phrase structure formalism of the Penn Treebank (PTB; Marcus et al. 1993). However, he also mentioned in passing that (approximately cited from my memory), much as he would have liked to use PTB as representing Present-Day English, in his view the annotation quality of the corpus was not good enough for his purposes. Relevant in this context is that PTB at that time was widely considered to be *the* gold standard treebank for English NLP, against which in particular phrase-structure parser accuracy was routinely measured (cf. Zaenen 2006). Kroch's picture is confirmed by a number of publications describing efforts to develop automatic or semi-automatic methods for finding annotation errors in corpora (e.g. Dickinson and Meurers 2003), although it is somewhat unfair to single out PTB in this way, when the truth is that it was used in these experiments primarily because of its wide availability and meticulous documentation, not because its annotation quality was believed to be inferior to that of other treebanks.

2. Annotation for gold standard NLP data

The fundamental assumption licensing the use of gold standards for evaluation in NLP is of course that the annotations in the gold standard are correct – that they constitute the *ground truth* in the domain in question – or at least that their degree of correctness is known.

This is where the – often invoked but frequently misunderstood – measure of *interannotator agreement* enters the picture.

2.1. Interannotator agreement in NLP

Interannotator agreement (IAA or ITA)[4] is a measure of annotation reliability (see Artstein and Poesio 2008). The measurement of IAA as originally formulated presupposes that certain preconditions are fulfilled (Artstein and Poesio 2008:574):

- There is more than one annotator
- There must be a detailed and clear annotation manual
- There must be clear explicit criteria for selecting annotators
- The annotators must work independently of each other

In NLP, much effort has been spent on devising fair measures of IAA. Most accounts of gold standard annotation found in the literature do not discuss annotation quality or accuracy. Instead, IAA is reported without comment as if it were such a measure.[5] That this is a mistaken belief is easily seen if we imagine a thought experiment where naive annotators are supplied with an explicit and clear, but false annotation manual.

IAA as practiced in NLP was originally developed for content coding (see Carletta 1996), and because of this annotation is seen as analogous to conducting a scientific experiment, with concomitant requirements of replicability, etc., hence the insistence on the preconditions listed above (Artstein and Poesio 2008:574). Notably, the use of expert annotators is sometimes explicitly eschewed, or at least frowned upon, since experts may make annotation decisions

[4] Also called *inter-coder agreement*.

[5] A notable exception to this is the first widely used NLP gold standard: The relationship of IAA to the quality of the part of speech annotation produced by the Penn Treebank annotators was estimated using the POS-tagged version of the Brown Corpus (Kucera and Francis 1967) as a gold standard (Marcus et al. 1993:318–320).

based on their domain expertise rather than on anything written in the annotation manual, thereby potentially compromising reproducibility. However, Artstein and Poesio (2008) and Artstein (2017) do note that annotation for NLP purposes often deviates from this methodology.

Some kinds of linguistic annotation are indeed similar to content coding (see further below), but "low-level" linguistic annotation such as for parts of speech, syntactic structure, discourse segments, coreference, and word senses, are arguably not among them. Instead, this kind of annotation is more akin to e.g. medical diagnosis, i.e., the remit of experts – highly trained professionals with long practical experience – rather than an activity which laypersons can engage in successfully on the basis of the contents of even very detailed annotation guidelines.

Importantly in this context, it is generally agreed that a key component of expertise is *intuition* (Hetmański 2018), a catch-all label summarizing a kind of Gestalt knowledge formed by long formal training and extensive practical experience. Thus, it may not be possible to have explicit and exhaustive annotation guidelines instead of expert annotators. In fact, even when employing non-experts for linguistic annotation, it seems to me that a lot of background knowledge about language description is assumed ("school grammar"), and not explicitly provided in an annotation manual.

2.2. Interannotator agreement and annotation quality

Given that gold standards have such pride of place in present-day NLP, it is somewhat surprising that studies into the methodology of gold standard compilation are few and far between, the exception being works dedicated to the formal aspects of interannotator agreement (see Artstein and Poesio 2008, Artstein 2017, and references provided there). Similarly, since practical experience teaches us that failure to report IAA may be seen as sufficient grounds for rejection of conference papers, we would expect more and more varied investigations of how IAA relates to gold annotation quality than we actually see in the literature.

There are some studies – not of annotation quality directly – but of how various characteristics of annotation tasks influence IAA. Bayerl and Paul (2011) present a meta-study where they attempt to tease out contributions to IAA from factors such as *domain*, *complexity of annotation scheme*, *language* (of the text), *number of annotators*, and notably *annotator training* and *domain expertise*. There are also some studies which have looked specifically at differences between non-expert and expert annotators (e.g. Snow et al. 2008, Gillick and Liu 2010, Munro et al. 2010, Plank et al. 2014), as well as some studies of other factors influencing annotation quality (e.g. Babarczy et al. 2006, Sampson and Babarczy 2008, Brown et al. 2010).

For the factors annotator training and domain expertise, the meta-study by Bayerl and Paul (2011) is hampered by an easily observable fact about accounts of gold standard annotation efforts, viz. that the relevant background of annotators is rarely specified and also subject to a "Chinese whispers/Telephone" game effect when cited in other sources. For example, Snow et al. (2008) refer to five kinds of annotations made by "experts", without clarifying the credentials of these experts, however. The publications cited by Snow et al. (2008) where these datasets are presented describe their annotators as "human annotators" (Dagan et al. 2006), "[undergraduate] students at the State University of New York at Oswego [...] native speakers of English" (Miller and Charles 1991), "linguistics students" (Pradhan et al. 2007), "Double blind annotation by two linguistically trained annotators was performed on corpus instances, with a third linguist adjudicating between inter-annotator differences" (Palmer et al. 2004:51), "annotators" (Strapparava and Mihalcea 2007), and "The initial stage was carried out by 5 annotators of remarkably different profiles with regards to their linguistic background. All of them however had participated in the development of the TimeML annotation scheme. The group of annotators for the second stage comprised 45 computer science undergraduate and graduate students" (Pustejovsky et al. 2003:652). Similarly, Hovy et al. (2014) refer to the annotators of the dataset described by Gimpel et al. (2011) as "experts" and "professional", while their original characterization by Gimpel et al. (2011) is simply as "researchers", where the inference is that these are researchers in a computer science department. They may of course have linguistic training, but we are not given any information about this.

2.3. The role of expertise in linguistic annotation for NLP

Some of the findings of the methodological studies cited above are, in brief summary:

- Teams of "experts" tend to show higher IAA than teams of "non-experts",
- but the differences are generally small.
- IAA is lower in mixed groups of experts and non-experts than in homogeneous groups (of both kinds).

- IAA is dependent on the task.
- IAA is dependent on the complexity of the annotation scheme.
- IAA is dependent on the number of annotators.

Most of these findings come from the meta-study by Bayerl and Paul (2011), where we learn many interesting things about linguistic annotation projects. However, new questions also arise in this connection which Bayerl and Paul (2011) do not address (presumably because of insufficient data). In particular, it would be useful to find out if training and expertise interact with other factors, for example, if the impact of annotation scheme complexity is different with experts and non-experts, if an increase in the number of annotators impacts IAA negatively to the same extent with experts and non-experts, etc.

Differences between expert and non-expert annotators have been noted by a number of authors (Kilgarriff 1999, Wilks 2000, Snow et al. 2008, Gillick and Liu 2010, Artstein 2017), and more generally for linguistic judgements by Dąbrowska (2010). Despite this, a general impression gleaned from accounts of NLP gold standard creation is that linguistic expertise acquired through formal academic training is undervalued.[6] It is difficult to understand the frequent omission of annotator qualifications in any other way, especially since we sometimes are told explicitly that the annotators are e.g. computer science students.

This much-discussed difference between expert and non-expert annotators in fact largely mirrors another, in my view more fundamental dichotomy in the kinds of annotations found in NLP gold standard datasets, to which we now turn.[7]

3. Baby and bathwater

A recent article by Uma et al. (2021) presents a useful summary of many issues in connection with IAA, and in particular in connection to the low IAA scores often reported for various kinds of NLP annotation tasks (e.g., by Lindahl et al. 2019 for argumentation annotation).[8]

They single out "inherently subjective judgments" as particularly amenable to variation in annotation results (and hence low IAA), but note that even annotation of "objective and 'simple' aspects of language" is not free from such problems (Uma et al. 2021:1387).

I believe that some of the discussion in the literature around annotation quality is confused by a mixup of two quite different kinds of annotation. By and large, the "objective and 'simple' aspects of language" mentioned above are facets of *linguistic analysis*, while the "inherently subjective judgments" are exactly this: judgements of (aspects of) language expressions made by language users. Filing both these activities under the heading "annotation" serves to obscure the fact that they are in reality quite different phenomena. The former requires (highly trained) experts, and the latter "only" ordinary language users. On a charitable interpretation, their conflation may be due to a belief that the native speaker's status as "expert" wielder of their language automatically also implies their expertise in formal linguistic analysis of the language, somewhat analogous to a belief that if you happen to inhabit and operate a human body, you will also automatically possess complete medical knowledge. This is a fallacy similar to the one discussed by Santana (2018) with regard to what should and should not count as scientific evidence.

Another reason for this ambiguity of the term "annotation" is surely to be found in the recent history of the NLP field. From the point of view of linguistics the current focus on deep-learning systems arguably represents a return to behaviorism (see, e.g., Alkon 1959, Passos and Matos 2007); annotation is framed as a black-box solution to a black-box problem. Instead of (scientific) analysis, annotation encodes (observational) data on which analyses are based.

This fits well with an observation made by Öhman (2021), that comparing lexicon-based and data-driven sentiment analysis is not comparing like with like. Word polarity values retrieved from a sentiment lexicon form a component in a suggested explanation of why a text or text passage will be perceived to carry a particular sentiment polarity, where the central explanatory device is tried-and-true linguistic compositionality. This may or may not be correct – this is an empirical matter – but it is emphatically not the same thing as attaching judgements about their sentiment to whole texts or text passages.

[6] How else are we to interpret statements such as "using experts is very expensive, prohibitively so for large-scale projects" (Uma et al. 2021:1386), with its implication that the experts are not really needed anyway.

[7] I am indebted to the two anonymous reviewers who both made remarks which steered my thinking in a – hopefully – fruitful direction.

[8] Uma et al. (2021) also propose and test concrete ways of compensating for low IAA in a machine learning setting, a topic which we will not be able to discuss further here.

From the point of view of somebody who still likes to believe that linguistics has something to contribute to annotation methodology in NLP, this wholesale return to behaviorism seems somewhat defeatist: the blanket label "inherently subjective judgments" is wielded in too cavalier a fashion, with sentiment analysis or detection of offensive language implicitly being put on a par with fashion preferences or individual (but in many cases shared) aversions to some words (see Liberman 2012). I suspect that similarly to how NLP research for a very long time largely ignored the quite mature linguistic subarea of language typology (Bender 2011; 2016), it seems that the field still remains unaware of the potential contributions by conversation analysis and text linguistics towards more objective analyses in these cases. This is not to deny that there are clearly more objective (or better: intersubjective) and more subjective language phenomena,[9] but I also believe that there is still a largely untapped resource (by NLP researchers) in the form of highly refined linguistic analysis of the "inherently subjective" phenomena (e.g. Klein 2018).

4. After the gold rush

Hopefully, it has become clear from the above, that to me, one of the more problematic aspects of NLP gold standard creation has to do with annotator qualifications and annotation quality.

I admit to being biased by having extensive linguistic training and having been exposed to large volumes of linguistically annotated text where IAA is unavailable for the simple reason that the annotation has invariably been carried out by only one expert (although drawing on native-speaker consultants). This is interlinear glossed text (IGT), resulting from linguistic fieldwork.[10]

Here are some methodological questions and hypotheses concerning the role of expertise in the form of formal training in linguistic analysis:

- Intuitively, we would expect different pairs of expert annotators to differ on more or less the same items, reflecting genuine disagreements of theory or practice (cf. Plank et al. 2014). This may mean that IAA will not be a very meaningful measure with expert annotators, and consequently that only one annotator will be needed in many cases (cf. IGT, mentioned above).
- Given the preceding point, what are the absolute limits to expert annotation in different domains (this is basically the question posed by Babarczy et al. 2006 and Sampson and Babarczy 2008)?
- Can we develop effective tools helping us to "clean up" inconsistent or erroneous annotations (cf. Dickinson and Meurers 2003, Dickinson 2009, Loftsson 2009, Kato and Matsubara 2010, Dickinson 2015, Hollenstein et al. 2016, de Marneffe et al. 2017)?
- The distinction between analysis and judgement is obviously primary when selecting an annotator pool. This will form a basis for when to solicit non-expert "products" and when an expert is required. However, this is not a strict dichotomy, since:
- Different linguistic analysis tasks seem differently tractable to lay annotators (Munro et al. 2010). Generally, it seems that annotation which focuses on analysis of linguistic form causes more difficulty to non-experts – except possibly (non-)existence of a particular form (such as a non-word) – than annotation focusing on (suitably granular semantic or pragmatic) content.
- Which kinds of linguistic training or expertise make a difference, and to which annotation tasks? Can we agree on how to specify annotator qualifications?[11]

5. A conclusion of sorts

Since this article has been written as a kind of methodological opinion piece, a definite conclusion is not so easily formulated. I can offer this, however: In order not to give in completely to the "new behaviorism" characterizing present-day NLP, it could be beneficial to ask annotators not only for judgements, but also for them to indicate what prompted (positive) judgements, similarly to McDonnell et al. (2016), who ask their annotators to provide (free-text) rationales

[9] There are of course also various kinds of linguistic – including idiolectal – variation, which may inform some kinds of annotation, but these are notably *not* subjective, but amenable to linguistic analysis.

[10] To be fair, I do find inconsistencies in such data, so that this form of annotation could benefit from consistency-checking NLP tools.

[11] It is at least questionable whether (even linguistics) student annotators can credibly be referred to as "experts".

for information-retrieval relevance assessment ratings. In a linguistic annotation context, annotation instructions could be, e.g., not "mark all instances of hate speech in these texts", but instead: "mark all instances of hate speech in these texts, together with the features (words, phrases, etc.) that in your view flag them as hate speech". Hence: a structural analysis task rather than a pure classification task.[12]

In addition, I have hopefully managed to convince the reader that there are many exciting methodological avenues to explore in the area of linguistic annotation for NLP, with the potential to contribute to improving the quality of gold standard datasets, or at least to improving our confidence in them.

Acknowledgements

I extend my heartfelt thanks to two anonymous reviewers for sharing useful literature references and for posing some pointed questions which have forced me to sharpen my thinking about the issues discussed here.

References

Alkon, Paul K. 1959. Behaviourism and linguistics: An historical note. *Language and Speech* 2 1: 37–51. https://doi.org/10.1177/002383095900200105.

Antonsen, Lene, Trond Trosterud, and Linda Wiechetek. 2010. Reusing grammatical resources for new languages. In *Proceedings of LREC 2010*, pp. 2782–2789. ELRA, Valletta.

Artstein, Ron. 2017. Inter-annotator agreement. In *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, pp. 297–313. Springer, Dordrecht. https://doi.org/10.1007/978-94-024-0881-2_11.

Artstein, Ron and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics* 34 4: 555–596. https://doi.org/10.1162/coli.07-034-R2.

Babarczy, Anna, John Carroll, and Geoffrey Sampson. 2006. Definitional, personal and mechanical constraints on part of speech annotation performance. *Natural Language Engineering* 12 1: 77–90. https://doi.org/10.1017/S1351324905003803.

Bayerl, Petra Saskia and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics* 37 4: 699–725. https://doi.org/10.1162/COLI_a_00074.

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6 3.

Bender, Emily M. 2016. Linguistic typology in natural language processing. *Linguistic Typology* 20 3: 645–660. https://doi.org/10.1515/lingty-2016-0035.

Brown, Susan Windisch, Travis Rood, and Martha Palmer. 2010. Number or nuance: Which factors restrict reliable word sense annotation? In *Proceedings of LREC 2010*, pp. 3237–3243. ELRA, Valletta.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* 22 2: 249–254.

Church, Kenneth Ward and Joel Hestness. 2019. A survey of 25 years of evaluation. *Natural Language Engineering* 25 6: 753–767. https://doi.org/10.1017/S1351324919000275.

Dąbrowska, Ewa. 2010. Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27: 1–23. https://doi.org/10.1515/tlir.2010.001.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, edited by Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché-Buc, pp. 177–190. Springer, Berlin.

Dickinson, Markus. 2009. Correcting dependency annotation errors. In *Proceedings EACL 2009*, pp. 193–201. ACL, Athens.

Dickinson, Markus. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass* 9 3: 119–138. https://doi.org/10.1111/lnc3.12129.

Dickinson, Markus and W. Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of EACL 2003*, pp. 107–114. ACL, Budapest.

Gillick, Dan and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 148–151. ACL, Los Angeles.

[12]This would also mesh nicely with a growing interest on the part of the NLP community in *explainable AI*, with work on "probing" neural networks for internal structures corresponding to conventionally assumed linguistic information (e.g., Şahin et al. 2020).

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of ACL/HLT 2011*, pp. 42–47. ACL, Portland.

Hämäläinen, Mika and Khalid Alnajjar. 2021. The Great Misalignment Problem in human evaluation of NLP methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 69–74. ACL, Online.

Harrigan, Atticus G., Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology* 27 4: 565–598. https://doi.org/10.1007/s11525-017-9315-x.

Hetmański, Marek. 2018. Expert knowledge: Its structure, functions and limits. *Studia Humana* 7 3: 11–20. https://doi.org/10.2478/sh-2018-0014.

Hollenstein, Nora, Nathan Schneider, and Bonnie Webber. 2016. Inconsistency detection in semantic annotation. In *Proceedings of LREC 2016*, pp. 3986–3990. ELRA, Portorož.

Hovy, Dirk, Barbara Plank, and Anders Søgaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of ACL 2014*, pp. 377–382. ACL, Baltimore. https://doi.org/10.3115/v1/P14-2062.

Kato, Yoshihide and Shigeki Matsubara. 2010. Correcting errors in a treebank based on synchronous tree substitution grammar. In *Proceedings of ACL 2010*, pp. 74–79. ACL, Uppsala.

Kilgarriff, Adam. 1999. 95% replicability for manual word sense tagging. In *Proceedings of EACL 1999*, pp. 277–278. ACL, Bergen.

Klein, Gabriella B. 2018. Applied linguistics to identify and contrast racist 'hate speech': Cases from the English and Italian language. *Applied Linguistics Research Journal* 2 3: 1–16. https://doi.org/10.14744/alrj.2018.36855.

Kucera, Henry and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence.

Liberman, Mark. 2012. Literary *moist* aversion. `https://languagelog.ldc.upenn.edu/nll/?p=4389`. Language Log post.

Lindahl, Anna, Lars Borin, and Jacobo Rouces. 2019. Towards assessing argumentation annotation – a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pp. 177–186. ACL, Florence. https://doi.org/10.18653/v1/W19-4520.

Loftsson, Hrafn. 2009. Correcting a POS-tagged corpus using three complementary methods. In *Proceedings of EACL 2009*, pp. 523–531. ACL, Athens.

Manning, Christopher D. 2015. Last words: Computational linguistics and deep learning. *Computational Linguistics* 41 4: 701–707. https://doi.org/10.1162/COLI_a_00239.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19 2: 313–330.

de Marneffe, Marie-Catherine, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the Universal Dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 108–115. LiUEP, Pisa.

McDonnell, Tyler, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pp. 139–148. AAAI Press, Palo Alto.

Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6 1: 1–28. https://doi.org/10.1080/01690969108406936.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 122–130. ACL, Los Angeles.

Öhman, Emily. 2021. The validity of lexicon-based emotion analysis in interdisciplinary research. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, pp. 7–12. ACL, Online.

Palmer, Martha, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In *Proceedings of ScaNaLU 2004 at HLT-NAACL 2004*, pp. 49–56. ACL, Boston.

Passos, Maria de Lourdes R. da F. and Maria Amelia Matos. 2007. The influence of Bloomfield's linguistics on Skinner. *Language and Speech* 30 2: 133–151.

Plank, Barbara, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of ACL 2014*, pp. 507–511. ACL, Baltimore. https://doi.org/10.3115/v1/P14-2083.

Pradhan, Sameer, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of SemEval 2007*, pp. 87–92. ACL, Prague.

Pustejovsky, James, Patrick Hanks, Roser Saur, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics 2003)*, pp. 647–656. Lancaster University, Lancaster.

Reiter, Ehud. 2007. Last words: The shrinking horizons of computational linguistics. *Computational Linguistics* 33 2: 283–287. https://doi.org/10.1162/coli.2007.33.2.283.

Şahin, Gözde Gül, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics* 46 2: 335–385. https://doi.org/10.1162/coli_a_00376.

Sampson, Geoffrey and Anna Babarczy. 2008. Definitional and structural constraints on structural annotation of English. *Natural Language Engineering* 14 4: 471–494. https://doi.org/10.1017/S1351324908004695.

Santana, Carlos. 2018. Why not all evidence is scientific evidence. *Episteme* 15 2: 209–227. https://doi.org/10.1017/epi.2017.3.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*, pp. 254–263. ACL, Honolulu.

Strapparava, Carlo and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of SemEval 2007*, pp. 70–74. ACL, Prague.

Trosterud, Trond. 2006. Grammatically based language technology for minority languages. In *Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Language Technology*, edited by Anju Saxena and Lars Borin, pp. 293–315. Mouton de Gruyter, Berlin.

Trosterud, Trond. 2012. A restricted freedom of choice: Linguistic diversity in the digital landscape. *Nordlyd* 39 2: 89–104.

Uma, Alexandra N., Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research* 72: 1385–1470.

Wilks, Yorick. 2000. Is word sense disambiguation just one more NLP task? *Computers and the Humanities* 34: 235–243.

Wintner, Shuly. 2009. What science underlies natural language engineering? *Computational Linguistics* 15 4: 641–644. https://doi.org/10.1162/coli.2009.35.4.35409.

Zaenen, Annie. 2006. Last words: Mark-up barking up the wrong tree. *Computational Linguistics* 32 4: 577–580.

The Preconceptual Basis of Noun Class (Gender)*

Patrik Bye
Nord University

Abstract

Noun class is widely seen as "standing out" from other morphosyntactic categories in having a basis in ontological beliefs, or a 'semantic core'. The consequence of this view is that noun classes in natural languages frequently do not cohere semantically. Here I motivate an aspectual alternative according to which noun class is grounded in low-level cognitive processes including the detection of agency and sex-related cues (including shape/size) and 'mode' of attention. This suggests a way of bringing noun class more into line with the perspectivizing contribution of morphosyntactic features in general.

Keywords: noun class, morphosyntactic features, attention, agency detection, sex discrimination

Twenty years ago there was a flurry of research in Tromsø on the principles governing gender assignment in natural language. Trond played a leading role in all of this activity, elaborating meticulous and thought-provoking analyses of the gender assignment rules of Norwegian and Old Norse (published as Trosterud 2001; 2006). I was little more than a spectator at the memorable gatherings where these ideas were hashed out, but they were nonetheless formative, leaving me with intriguing questions to ponder in the ensuing years about human language, culture and mind. It was always my intention to address some of them in writing but, despite a couple of later workshop presentations, I let the opportunity slide. The occasion of Trond's sixtieth birthday is thus a fine opportunity to make a reconnection with some of these questions—and those exciting days in Tromsø at the turn of the millennium.

1. Noun class assignment: belief vs. perception

The prevailing view of noun class (gender) is that it has a 'semantic core' (Corbett 1991), which generally involves one or more of the dimensions of animacy, humanness and sex (Aikhenvald 2000:22).[1] However, this same idea underlies a widespread perception that, "[i]f we compare gender with the other morphosyntactic features, it seems evident that gender stands out" (Corbett 2014:87).

This belief-based view of noun class immediately faces two embarrassments. The first is the existence of 'partially semantic' systems that admit of apparently arbitrary assignments not obviously motivated in terms of the semantic core. The second is that, in some languages, nouns may be assigned to classes based on their morphological or phonological properties (formal assignment rules). The focus here is the problem of the apparent semantic incoherence of some noun classes. The formal assignment problem is beyond the scope of the present short article, but I will return briefly to the question at the end of the paper.

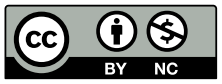
Both types of anomaly may be seen as artefacts of the belief-based approach to noun class. This approach has nevertheless been extremely fruitful, illuminating the distribution of a feature that was once believed to be largely lexically determined in terms of assignment rules of impressive accuracy and coverage. The work of Steinmetz (1986; 2006) and Trosterud (2001; 2006) exemplify this kind of approach to the complex systems of the Germanic languages.

The other way to 'limit the arbitrary', to borrow a phrase from Joseph (2000), is by trying to uncover a coherent natural basis for the pattern. While the motivation seems clear enough in the case of 'strict semantic systems', noun classes in 'partially semantic systems' may be made up of nouns assigned by rules

*I would like to thank an anonymous reviewer and Laura Janda (who relinquished her anonymity as reviewer for this piece) for their helpful comments. The core ideas outlined here have been presented on previous occasions at workshops at UiT in honour of Peter Jurgec's dissertation defence (14 January, 2011), and at the Center for Practical Knowledge seminar series at Nord (21 March, 2013). I would like to thank participants on those occasions, in particular Curt Rice and James McGuirk for their comments.

[1] I follow the growing trend of using *noun class* rather than *gender* to refer to the morphosyntactic category. However, if the ideas sketched here are on the right track, neither term is especially apposite. See 'Conclusions' for discussion of a possible alternative.

 Nordlyd 46.1: 27-36, *Morfologi, målstrev og maskinar: Trond Trosterud fyller | täyttää | deavdá | turns 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6371



that do not cohere semantically. The beginnings of a solution, I suggest, is to shift perspective away from *classification* to what Seiler (1986) calls *apprehension*, and instead view noun class assignment as rooted in lower order cognition, specifically the processes that govern the direction, fixation and focus of attention. How these processes interact with language structure is also the focus of much work in cognitive linguistics (see, e. g., Langacker 2008). However, I shall assume that these processes are governed by biases towards environmental cues of certain types, and it is these that can be discerned in patterns of noun class assignment in natural language.

The idea that noun class is perspectival is not new, although its application has so far been limited, and not terribly well known. The proposal nonetheless goes back to Brugmann (1897), whose ideas were later elaborated by Lehmann (1958) and Weber (1999), who explicitly argues that noun class can be viewed as nominal aspect (cf. Rijkhoff 1991), that is, the distinction between singulatives, collectives and 'continuatives' (mass nouns). More recently, Wiltschko (2012) has proposed to analyse the animate/inanimate distinction in Alqonquian as an expression of nominal aspect.

In what follows, we will examine noun class assignment as an expression of the detection of agency (Section 2), humanness and masculinity/femininity (Section 3), and composition (Section 4). Section 5 then introduces the idea, based on Mylne (1995), that noun classes in some languages may be assigned on the basis of what we can call 'mode' of attention. Finally, Section 6 suggests how some of these ideas might be followed up in future work.

2. Agency

A noun class distinction between 'animate' and 'inanimate' is found in several unrelated languages. Corbett (1991:20ff.) includes a lengthy discussion of the assignment to animate and inanimate gender in Ojibwe (oji; Algic, Algonquian; United States, Minnesota, Wisconsin, North Dakota, Montana/Canada, Ontario, Manitoba, Saskatchewan). Nouns denoting humans, animals, spirits, and trees are animate in this language, with the residue being inanimate. The complication is that some inanimate nouns are promoted to animate status, although the reverse does not occur. Examples of such 'promoted' nouns include *a:kim* 'snowshoe', *meskomin* 'raspberry', and *uppwa:kan* 'pipe' (for smoking). Black-Rogers (1982) makes the case that the apparent anomaly of many such assignments evaporates once we factor in the Ojibwe worldview, which posits an omnipresent life force that flows through all things to a greater or lesser extent. Promotion of an inanimate to animate noun class reflects the judgment that the thing in question is a focal point for this force.

Although it may be attractive to explain the anomalies of Ojibwe noun class assignment with reference to the belief system, this asymmetrical tendency to promote inanimates—what Hockett (1966) terms the 'absorptive' property of the animate noun class—is not unique to Algonquian, and therefore stands in need of explanation. We can find similar patterns when we look at unrelated languages with a similar animate/inanimate distinction. One such example is Car (caq; Austroasiatic, Nicobarese; India, Nicobar Islands Braine 1970), where the animate class includes inanimate nouns relating to human activity and motion, such as *á p* 'canoe', *c´ cə* 'surfboard', *sakú* 'knife', *linr̃´ñ* 'bow', *c´ k* 'arrow', *p´nsī l* 'pencil'.

What this 'absorption' brings to mind is Dawkins' (2006) proposed 'hyperactive agency detection device', or 'HADD', to which he imputes the cross-cultural tendency for humans to form beliefs in supernatural agencies. The hyperactivity consists in a bias towards detecting agency even where none is objectively present, potentially generating false positives. Such a bias is adaptive: individuals whose devices were less hyperactive would more often fail to recognize potential threats in the environment due to predators and human hostiles, with negative consequences for reproductive fitness. Natural selection has therefore honed an agency detection device with hyperactive properties.[2] If one by-product of such a device is supernatural beliefs, another by-product may be linguistic, consisting in a certain tendency for animate noun class to

[2]It is relevant in this connection to mention the proposal of Tichy (1993), who argues that Proto-Indo-European noun class was predicated on the marking of agent-patient distinctions. According to Tichy, Proto-Indo-European can be reconstructed as having had two genders, a *distinctum* which marked a contrast between agent and patient, and *indistinctum* which did not.

'absorb' inanimate nouns referring to things that may evidence agency.

The appeal in linguistic treatments of anomalous noun class assignment to 'mythology' or 'worldview' does not ultimately explain the pattern. Noun class assignment and worldview are in fact both *explananda*, to be explained by some deeper cognitive principle—such as agency detection. The reason that trees are assigned to the animate gender in Ojibwe may ultimately have less to with the occult 'power' that Black-Rogers sees as underlying the Algonquian worldview than the fact that movement in trees can activate the HADD. The role of culture in this explanation is to take these systematic features of experience and reinforce them, for example by constructing them as salient manifestations of 'power', or by lexicalizing the noun class of certain inanimate nouns as 'animate'. Rustling in the trees may be an inorganic consequence of the wind, but noun class assignment is primarily about how a referent might impact attention, and only secondarily about worldview.

Such an approach might be the beginning of an explanation of some of the more idiosyncratic assignments as well. To use a famous example from the related language Northern Cheyenne (chy; Algic, Plans Algonquian; US, Montana, Oklahoma), the fact that 'raspberry', but not 'strawberry', is animate, should not be taken to mean that speakers *conceive* of raspberries as animate (cf. Corbett 1991:23). If I were to hazard a guess at what motivates this particular assignment, it might be the fact that the raspberry bush is tall and tree-like, and may therefore be expected to trigger the HADD with greater frequency than the low-lying strawberry plant.

This account shifts the focus away from the ontologically-based distinction between animate and inanimate and onto the detection of agency, which is upstream of beliefs about the things of the world, and how this pre-conceptual engagement with the world is organized in experience.

3. Masculinity/femininity

If the HADD is one automatic detection system in the human repository, there are at least two others which may turn out to be important for understanding the function of noun class and how it is assigned. These are the ability to recognize other humans, most importantly through face perception, and the detection of masculine and feminine traits, as distinct from the brute fact of biological sex that is generally invoked as having semantic core function. A suitable collective term for masculinity and femininity would rather seem to be lacking. The term *gender* itself would have fit the bill well enough had it not been for its long history in linguistics, as well as its more recent usage in approaches that emphasise the constructed and performative aspects of gender as a social marker (see, e. g., Talbot 2019 [1998]). In the absence of a convenient term, I shall simply use the disjunction 'masculinity/femininity' rather than 'sex'.

Face perception is notoriously hyperactive, as demonstrated by the universal predisposition for seeing faces in clouds, rocks, trees, and so forth (e. g., Guthrie 1993). If the detection of characteristics associated with sex is similarly hyperactive, this could potentially explain how parameters such as size and shape in inanimates and non-sex-differentiable animates may condition assignment to masculine or feminine noun class without having to appeal to higher-level cognition, including beliefs and ideologies. Since height, body size, and fat distribution are important sexually dimorphic cues to health, and therefore mate choice (Sugiyama 2016), we might expect to see attention to these features manifested in noun class systems. Moreover, if attention to these features is activated in lower order perception, we would not necessarily expect an alignment of noun classes with ontological categories.

What Corbett (1991:8) calls "strict semantic systems" may be based on experiences where the detection of cues for humanness and masculinity/femininity act in concert. The paradigm example of such a system is Tamil (tam; Dravidian, Southern; India, Tamil Nadu/Sri Lanka, Eastern Province, Northern Province; Asher 1989 [1985]). In this language, the masculine/feminine distinction is reserved for 'rational' referents, which includes humans and divine beings, as well as a few inanimate nouns with a metonymic relation to the latter, such as *cuuriya* 'sun' and *cantira* 'moon', both of which are also designations for the associated male gods.

Given the assumption of a 'semantic core', it is surprising that 'strict semantic systems' seem to occur rather rarely compared with 'partially semantic systems'. If, on the other hand, noun class assignment in such systems presupposes the integration of lower order perceptual input from more than one detection system, this might help explain why they are the exception rather than the rule.

Another language where the distinction between masculine and feminine is ontologically transparent is Burushaski (bsk; isolate; Pakistan, Gilgit-Baltistan; Berger 1974, Munshi 2018). Burushaski has a system of four noun classes predicated on a fundamental distinction between human and non-human. We shall return to the non-human genders in Section 4. Only human referents are distinguished as masculine or feminine, e. g., *badśá* 'king (m)', *axón* 'priest (m)', *náni* 'mother (f)'.

In other languages, the masculine/feminine distinction is projected further down the animacy hierarchy, including to non-sex-differentiable animates and inanimates. By way of an example, consider Mian (mpt; Trans New Guinea, Ok; Papua New Guinea, Sandaun province, Telefomin district; Fedden 2011), which has a four-way system that provides an interesting contrast with Burushaski. In common with the latter, Mian also has masculine and feminine noun classes. However, included in the masculine and feminine classes of Mian are nouns for animals whose sex is "not readily discernible or relevant", but which are assigned "conventionalized gender". The main criterion for feminine assignment is shape, with many animals of squat or round shape assigned to the feminine class. Thus, the eagle (*tolim*) is masculine, while the cassowary (*koból*) is feminine.

As Aikhenvald (2000) shows, certain physical parameters including size and shape, as well as position and solidity, recur in the assignment of inanimates to masculine or feminine gender. This is probably no accident. The underlying reason that size and shape figure in such promotion should be sought in the automatic processes that feed the perception of body dimorphism in humans. Another relevant parameter is the presence of neotenic traits, which have been sexually selected for in humans, but in women in particular, and which elicit subjective perceptions of 'cuteness' and associated caretaking behaviours. For example, Dizi (mdx; Afro-Asiatic, Omotic; Ethiopia; Allan 1976) distinguishes a feminine and non-feminine gender, but also assigns 'cute' animals to the feminine, irrespective of biological sex.

The correlation of size and shape with masculinity/femininity can be illustrated with two examples. In Maasai (mas; Eastern Nilotic; Kenya/Tanzania; Payne 1998), the masculine and feminine noun classes are also productively used to convey augmentative or diminutive meaning. There is a core of nouns whose class is lexically fixed, but the class of most other nouns is determined by pragmatic context. The speaker's denigration of sexed referents picked out by gender-specific nouns may be signalled by using the opposite gender prefix, thus *en-títo* 'girl' vs. *ol-títo* 'large shapeless hulk of a woman', *l-ám`y`* 'male donkey' vs. *nk-ám`y`* 'wimpy male donkey'. With gender-neutral roots referring to sex-differentiable referents, pejorative connotations may also accompany a change in the sex of the referent, e. g., *en-kitók* 'woman' vs. *ol-kitók* 'very respected man', *l-abáánì* 'male doctor' vs. *nk-abáánì* 'female or small doctor, quack'. With inanimates, only size is relevant, e. g., *l-ál`m* 'sword' vs. *nk-ál`m* 'knife'.

The Cantabrian (Montañés) variety of Asturleonese (ast; Indo-European, Romance; Cantabria, northwest Spain; Holmquist 1991) has a similar shape-based noun class system in which the primary distinction is between referents perspectivized as narrow ('masculine', ending in *-u*) or wide ('feminine', ending in *-a*), e. g., *calleja* 'alley' vs. *calleju* 'narrow alley', *poza* 'quagmire' vs. *pozu* 'drinking well', *ría* 'valley' vs. *ríu* 'mountain river'. In these examples it is the 'feminine' term that designates the larger entity. In the case of terms referring to trees and their fruit, which form minimal pairs distinguished only by their noun class, this relation is apparently reversed, e. g., *cereza* 'cherry' vs. *cerezu* 'cherry tree', *manzana* 'apple' vs. *manzanu* 'apple tree', *panoja* 'ear of corn' vs. *panoju* 'cornstalk'. However, this is explained by the shape criterion: the 'masculine' form of the designations for trees foregrounds their height. In contrast to Maasai, it is the masculine that is most strongly correlated with pejorative readings of 'meagreness' (p. 68), e. g., *carretera* 'highway' vs. *carreteru* 'narrow, bumpy roadway', *oveja* 'sheep' vs. *oveju* 'sheep (meagre fare)'.

Although the relation between a referent's shape and the masculine/feminine distinction recurs across unrelated languages, no satisfactory explanation for it has yet been put forward. Grimm (1989 [1890]:343) writes in an oft-cited passage that "Das grammatische genus ist demnach eine in der phantasie der men-

schlichen sprache entsprungene ausdehnung des natürlichen auf alle und jede gegenstände". The key words here are *Phantasie* and *Ausdehnung*, 'extension', because they suggest that the explanation lies with higher-level cognitive processes, in particular the projection, by way of conceptual metaphors, of sexual characteristics onto entities that objectively lack them (e. g., Trosterud 2001). If, on the other hand, noun class assignment is grounded in lower level cognitive processes as I have suggested, it is not necessary to assume that 'worldview' is somehow inscribed there.

4. Composition

We now turn to features relevant in discriminating between classes of inanimate that do not involve promotion animate or masculine/feminine class. Some languages have noun classes assigned on the basis of quite specific cues. Anindilyakwa (aoi; Macro-Gunwinyguan, East Arnhem; Australia, Northern Territory, Groote Eylandt/Bickerton Island; Leeding 1989), for example, makes a distinction in non-personified nouns between visible and invisible and, within the visible class, between lustrous and lustreless. Ngan'gi (nam; Daly; Australia, Northern Territory, Daly River Region; Tryon 1974, Reid 2011 [1990]) has noun classes specifically for hunting weapons and anything made of wood.

However, the most common distinction is based on the perceived composition of the referent, which is related to the ontological distinction between individual and substance. Rijkhoff (1991) proposed a four-way distinction based on the features [structure] and [shape], yielding a contrast between collective nouns [+structure, +shape], mass nouns [+structure, –shape], individual nouns [–structure, +shape], and concept nouns [–structure, –shape]. Although widely relevant for the syntax of NPs, similar distinctions are also the basis for assignment to noun class in some languages, such as Burushaski, which has two neuter genders traditionally designated 'x' and 'y'. The former includes animates and tangible inanimates, while the latter includes non-individuated terms: abstracta, aggregates and substances. Trees are viewed as aggregates, thus *branç* meaning 'mulberry' is x, but in the sense 'mulberry tree' is y.

Once again, Mian provides an interesting contrast with Burushaski. Like Burushaski, Mian has two neuter noun classes (1 and 2) for inanimates (Fedden 2011:174f.). In addition to body parts (*bān* 'arm'), natural entities (*deit* 'bird's nest') and cultural artefacts (*was* 'drum'), neuter 1 also includes liquids and substances (*deib* 'moss', *isá* 'pus'). Neuter 2 is used with what Fedden designates 'masses' (*awitnîn* 'stars'), locations and landmarks (*kwoisâm* 'spirit house'), weather phenomena (*íb* 'clouds'), illnesses (*kweim* 'fever'), intangibles and abstract nouns (*fotom* 'shame'), temporal and verbal nouns (p. 175). While Burushaski perspectivizes the distinction between individuated (x) and non-individuated (y), it can appear that the basis of the Mian system is part-whole focus. Thus, neuter 2 nouns are often aggregates or superordinates, such as *afobèing* 'goods', *fub* 'rubbish bits', *kibi* 'face (the collective of eyes, nose, mouth)', and *unín* 'food'.

As Aikhenvald (2000) points out, 'solidity' is also one of the recurring parameters involved in assigning masculine or feminine gender. Lehmann (1958) proposed that the three-way gender system of Proto-Indo-European was predicated on composition rather than biological sex. The 'feminine' rather served to mark a noun as collective, while the 'masculine' had a singulative function. Thus, an individual cold or frost was masculine *himá-s*, while the feminine *hima-h* referred to a sequence of such events, that is, winter. The neuter *hima-m* was according to Lehmann a resultative form meaning 'snow', but could equally be a mass noun. Such a system may have survived into early Germanic, as Leiss (1999) argues was the case for Old High German on the basis of the large number of nouns with double or triple gender attestations. If this is correct, it raises interesting questions about the relation between composition-based assignment and assignment based on perceived agency and masculinity/femininity.

A similar question is raised by Tayap (gpn; unclassified; Papua New Guinea, East Sepik Province), which Kulick and Terrill (2019:58) claim assigns gender in the following way. If a noun is particularized as opposed to generic, it is masculine by default, otherwise feminine. If particularized and has a non-male referent, it assigned to the feminine class. Particularized male referents receive masculine gender if long, but feminine if 'stocky'.

5. Mode of attention

In the preceding sections we have proposed an approach to major noun class parameters in terms of apprehension rather than ontology of the referent. These parameters ultimately involve the detection of agency, humanness, masculinity/femininity and composition.

In some languages, however, the assignment of inanimates to masculine and feminine noun classes does not appear motivated by any of the features discussed so far. A relevant proposal by Mylne (1995) reanalyses the masculine-feminine distinction in Dyirbal (dbl; Pama-Nyungan; Australia, Northeast Queensland; Dixon 1972) in terms that do not directly have to do with the detection of masculine/feminine traits. What instead appears to distinguish the relevant classes is what we can call a 'mode' or quality of attention.

Dyirbal has four noun classes, diagnosed by the choice of classifier that accompanies a noun in a noun phrase. According to Dixon's (1972:308) simple schema for class membership, the classifier *bayi* is used with (human) males and non-human animates, *balan* for (human) females, water, fire, and lightning. The *balam* class includes 'non-flesh food' (edible vegetables, fruit and honey), and the residue *bala* class contains everything not in the others.

Exceptions to these generalizations are accounted for through three principles. The myth-and-belief principle (cf. the discussion of Ojibwe above) explains why birds, which in mythology represent dead human females, are assigned to the *balan* rather than the *bayi* class. The second is metonymy. Fishing implements might be expected to be *bala*, but are exceptionally *bayi*, apparently by association to fish. The third is that things that have the capacity to cause harm are assigned to *balan*. For example, fish are in general *bayi* by virtue of being animate, but the stone fish and gar fish are *balan*.

The question is what connects harmful things to a noun class whose semantic core is allegedly female sex. Lakoff (1987:92–104) analyses the *balan* class as a radial category in which the concepts relating to women, fire and dangerous things are 'chained' by experiential links. The reason that the sun is assigned *balan* is mythological—the moon is the 'husband' of the sun. This permits the attraction of other nouns to the *balan* class. Since the sun and fire are from the same 'domain of experience', fire is also assigned to *balan*, from where it attracts other 'dangerous' things, including stinging nettles, fighting spears, gar fish, and water.

In seeking to restore the *balan* class as structured on a single ICM, Mylne (1995) criticizes the tenuousness of Lakoff's links and argues his approach entails imposing Western categories of thought on the system. Mylne's point of departure is Dixon's observation that the *bayi* class is based on animacy rather than maleness, and casts doubt on the idea that the distinction between *bayi* and *balan* is ultimately based on sex at all, which is a core assumption of Lakoff's account. Although Mylne leaves important questions unanswered, I nevertheless think it is the right direction to take.

At the core of Mylne's proposal is his observation that members of the *balan* class are frequently the source of disharmony or 'trouble', making *balan* a 'handle-with-care' tag. Moreover, he claims (p. 394): "it is the non-obvious capacity to cause trouble and the existence of a cultural reason for avoidance which seem to be relevant to the *balan* class". Elaborating with more specific examples, he goes on:

> Stonefish are extremely dangerous, the more so because of their excellent camouflage. The platypus is small, attractive, and apparently harmless, but inflicts a vicious wound when handled. The same applies to the hairy mary grub. Stinging trees and stinging nettles inflict pain on the unwary, but their capacity to do so is not self-evident; one must learn to recognize them. Fighting spears, shields and fighting grounds are sources of trouble, but those who do not know this need to be taught it; children need to be warned away. […] Fire is clearly a potential source of trouble […], but again, this may not be obvious to the inexperienced. Water bodies are home to dangerous and unpredictable spirits (the mythological associations probably take precedence over, but would be associated with, the risk of being attacked by crocodile [sic!] or of drowning).
>
> (Mylne 1995:387f.)

Thus, *balan* marks the referent of a noun as requiring a certain circumspection. As Mylne himself writes (pp. 390–393), this raises the question why women rather than men should be assigned to this class.

It can be explained, at least partly according to Mylne, with reference to the avoidance behaviours required by Dyirbal culture. In addition to an everyday form of the language known as 'Guwal', there was a separate lexicon, Dyalŋuy, or 'mother-in-law language', which had to be used in the presence of taboo relatives. Dyalŋuy remained in use until about the 1930s.

According to Dixon (1972:32), "the rules for using Dyalŋuy […] precisely indicate who is sexually available for any person." These rules were acquired young, since "[c]hildren were promised in marriage at an early age, thus acquiring a full set of taboo relatives; Dyalŋuy was probably learned in the same way as Guwal, perhaps a year or two behind it." The penalty for failing to observe the taboos ranged from being publicly shamed to being put to death.

This explanation cannot be complete, however, since the taboo was reciprocal: the use of Dyalŋuy was obligatory in the presence of parents-in-law, children-in-law, and cross-cousins (father's sister's or mother's brother's child) irrespective of ego's own sex. This suggests there must be some additional factor at work that explains the assignment of women to *balan*. The styling of Dyalŋuy specifically as 'mother-in-law language' also points to an asymmetry behind the ostensibly symmetrical taboo.

One possible explanation is cultural, but Mylne rejects the idea that women are regarded as intrinsically troublesome in Dyirbal society, pointing to the equally treacherous roles that men and women play in Dyirbal myth. He hypothesizes instead (p. 392) that, underlying 'femaleness' and 'trouble' is a model of "the other, the extra-ordinary, that which is set apart as being associated with the potential to disrupt harmony". This formulation does not appear to me, however, to eliminate the fundamental disjunction in a way that would allow Mylne to restore a single idealized cognitive model (ICM) for the *balan* class. One of the alternatives that Mylne considers, however, is that "it is women who impose the burden of avoidance on men rather than vice versa". Opting for this route would seem to lead to the conclusion that gender asymmetries are inscribed in the linguistic system itself: "That the language of the community as a whole should embody such a belief implies a dominance of the male point of view which is certainly found in western language and culture." Although Mylne does not choose this option, Nesset (2001) makes a parallel argument for Russian on the basis of declension class distribution. However, such a conclusion would be premature. A reasonable alternative would be to seek a basis for the asymmetry beyond the domain of culture.

Buss (2019 [1998]:159–186) lays out key evidence for the existence of evolved asymmetries between male and female short-term mating strategies that show a high degree of stability across cultures. In particular, men are more strongly motivated on average to seek out sexual variety. This in turn predicts that they are more likely to show an interest in women that are culturally off-limits, and risk the penalties for doing so. If this is on the right track, it can help makes sense of why women should be marked linguistically as requiring circumspection—without appealing to mythologically or ideologically inscribed beliefs.

I also think that we can put to rest Mylne's concern that such an account would imply "dominance of the male point of view", since the costs of failing to exercise proper circumspection are typically not incurred individually, but severally—even community-wide—irrespective of gender. The appropriate circumspection required by something marked as *balan* can be understood as a shared responsibility. Straying too close to the water's edge and being attacked and maimed by a crocodile is a cost borne not just by the victim, but by the victim's family, and so it therefore behooves everyone to be on their guard. In the same way, the costs of being caught making eyes at a taboo relative fall not just on the would-be lover, but also kin and affinal relatives, of either sex, whose lives would be threatened with disruption should an affair come to pass.

Before closing this section, it is important to address an objection against both Lakoff's and Mylne's analyses raised by Plaster and Polinsky (2007) on learnability grounds, which could be extended to the claims advanced here. I will briefly explain why I think the objection does not apply. They write:

> A child has no inherent (or learned) association of women with dangerous things, as Lakoff argues, or as "other" and "associated with the disruption of the harmony of living", as Mylne (1995:387) proposes. Since many of the concepts that Lakoff and Mylne identify as underlying the Dyirbal noun class system are beyond the scope of young children's understanding, the systems posited by Lakoff and Mylne would be nearly impossible for children to learn.
>
> (Plaster and Polinsky 2007:19)

The consequence of this claim is that they give up on the semantic coherence of the *balan* class, positing separate statements assigning *balan* to nouns with different semantic labels, including [female], [fresh water], [fire] and [stinging].

I agree with Plaster and Polinsky's criticism of Lakoff, but their criticism of Mylne seems slightly misplaced, even if he did not ultimately resolve the disjunction between 'femaleness' and 'trouble'. Although children may not have the "concept" that something may be "associated with the disruption of the harmony of living", children are socialized into a habitus from when they are born, and this habitus includes the acquisition of avoidance behaviours from those around them (cf. Bourdieu 2000 [1972], Dreyfus 1991). There is no need to assume that the capacity to cause harm has to be perceived in the object that causes it, for example, by actually getting stung by a nettle. It is sufficient that this capacity is reflected in the circumspection of others in the community towards it. The same is true of the semantically much simpler *balam* 'non-flesh food' class. The main evidence that something is edible is that others eat it without spitting it out or falling sick.

In sum, the implication of Mylne's analysis is that sex (or masculinity/femininity) is not ultimately what underlies the distinction between the *bayi* and *balan* classes. It is rather something functional, encoding a difference in the way a noun's referent is generally attended to by Dyirbal speakers. It is no less perspectival for being lexically largely fixed. Although Mylne frames his account in terms of Aboriginal 'worldview', it actually furnishes a paradigm for grounding noun class assignment in cognitive processes upstream of belief formation. In this way, it is compatible with a broader perspectival approach to morphosyntactic features.

6. Conclusions and outlook

The received view of noun class as having a 'semantic core' is based on the assumption that noun class assignment is grounded in ontological beliefs. While this approach has led to a greatly improved understanding of the generalizations underlying noun class assignment, it inevitably also leads to the view that the content of noun class as a feature is arbitrary, since certain assignments will be semantically anomalous or based on form rather than semantics.

I have proposed instead that noun class is grounded in human attention systems, specifically, the detection of cues relevant to the perception of agency, humanness, masculinity/femininity and composition, as well as what I have termed 'mode' of attention. Much of the reason that we find that inanimates may be promoted to classes based on these criteria, I suggest, has to do with the 'hyperactivity' of the underlying detection systems. Since they are low-level cognitive processes, we do not have to invoke beliefs about the referent, making it possible to bring noun class into line with the perspectival nature of other morphosyntactic categories. In short, it becomes less evident that "gender stands out".

If the ultimate basis of assignment to animate noun class is low-level agency detection rather than higher level beliefs, the linguistic study of noun class may be better served by terms other than *gender* or *class*, which are permeated with classificatory assumptions. If it were up to me, I might venture the term *handle*.

The approach sketched here may also open a possibility, at least in principle, of explaining how noun class assignment might draw on formal criteria in some languages. Such a situation would be unexpected if noun class genuinely were grounded in ontology, but easier to reconcile with a grounding in attention/apprehension. It is perhaps relevant in this connection that certain kinds of sound symbolism are involved in making some of the same distinctions of size, shape and affective meaning as correlate cross-linguistically with the masculine/feminine noun class distinction (Ohala 1984).

I leave a fuller elaboration of these ideas and any broader implications for gender in language to future work.

References

Aikhenvald, Alexandra Y. 2000. *Classifiers: a typology of noun categorization devices*. Oxford University Press, Oxford.

Allan, Edward. 1976. Dizi. In *The non-Semitic languages of Ethiopia*, edited by M. Lionel Bender, pp. 377–392. African Studies Center, Michigan State University, East Lansing, MI.

Asher, R. E. 1989 [1985]. *Tamil*. Routledge, London.

Berger, Hermann. 1974. *Das Yasin-Burushaski (Werchikwar). Grammatik, Texte, Wörterbuch*. Otto Harrassowitz, Wiesbaden.

Black-Rogers, Mary B. 1982. Algonquian gender revisited: animate nouns and Ojibwa 'power'—an impasse. *Papers in Linguistics* 15 1: 59–76. https://doi.org/10.1080/08351818209370560.

Bourdieu, Pierre. 2000 [1972]. *Esquisse d'une théorie de la pratique*. Seuil, Paris.

Braine, Jean Critchfield. 1970. *Nicobarese grammar (Car dialect)*. Ph.D. thesis, University of California at Berkeley.

Brugmann, Karl. 1897. *The nature and origin of the noun genders in the Indo-European languages*. Charles Scribner's Sons, New York.

Buss, David M. 2019 [1998]. *Evolutionary psychology: the new science of the mind*. Routledge, London, 6th edn.

Corbett, Greville G. 1991. *Gender*. Cambridge University Press, Cambridge.

Corbett, Greville G. 2014. Gender typology. In *The expression of gender*, edited by Greville G. Corbett, pp. 87–130. Mouton de Gruyter, Berlin.

Dawkins, Richard. 2006. *The God delusion*. Transworld Publishers, London.

Dixon, R. M. W. 1972. *The Dyirbal language of North Queensland*. Cambridge University Press, Cambridge.

Dreyfus, Hubert. 1991. *Being-in-the-World: a commentary on Heidegger's Being and Time*. Van Gorcum, Assen, The Netherlands.

Fedden, Sebastian. 2011. *A grammar of Mian*. Mouton de Gruyter, Berlin.

Grimm, Jacob. 1989 [1890]. *Deutsche Grammatik 3. [= Jacob Grimm und Wilhelm Grimm. Werke. Forschungsausgabe. Abteilung I. Die Werke Jacob Grimms 12]*. Olms-Weidmann, Hildesheim.

Guthrie, Stewart. 1993. *Faces in the clouds: a new theory of religion*. Oxford University Press, Oxford.

Hockett, Charles F. 1966. What Algonquian is really like. *International Journal of American Linguistics* 32: 59–73.

Holmquist, Jonathan Carl. 1991. Semantic features and gender dynamics in Cantabrian Spanish. *Anthropological Linguistics* 33 1: 57–80.

Joseph, John E. 2000. *Limiting the arbitrary*. John Benjamins, Amsterdam.

Kulick, Don and Angela Terrill. 2019. *A grammar and dictionary of Tayap*. Mouton de Gruyter, Berlin.

Lakoff, George. 1987. *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago University Press, Chicago.

Langacker, Ronald W. 2008. *Cognitive grammar: a basic introduction*. Oxford University Press, Oxford.

Leeding, Velma J. 1989. *Anindilyakwa phonology and morphology*. Ph.D. thesis, University of Sydney.

Lehmann, Winfred P. 1958. On earlier stages of the Indo-European nominal inflection. *Language* 34: 179–202. https://doi.org/10.2307/410822.

Leiss, Elisabeth. 1999. Gender in Old High German. In *Gender in grammar and cognition, Vol. I: Approaches to gender*, edited by Barbara Unterbeck and Matti Rissanen, pp. 237–258. Mouton de Gruyter, Berlin.

Munshi, Sadaf. 2018. *Srinagar Burushaski: a descriptive and comparative account with analyzed texts*. Brill, Leiden.

Mylne, Tom. 1995. Grammatical category and world view: Western colonization of the Dyirbal language. *Cognitive Linguistics* 6: 379–404. https://doi.org/10.1515/cogl.1995.6.4.379.

Nesset, Tore. 2001. How pervasive are sexist ideologies in grammar? In *Language and Ideology: Cognitive Theoretical Approaches*, edited by R. Dirven, E. Sandikcioglu, and B. Hawkins, pp. 197–227. John Benjamins, Amsterdam.

Ohala, John J. 1984. An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica* 41: 1–16. https://doi.org/10.1159/000261706.

Payne, Doris L. 1998. Maasai gender in typological perspective. *Studies in African Linguistics* 27 2: 159–175. https://doi.org/10.32473/sal.v27i2.107385.

Plaster, Keith and Maria Polinsky. 2007. Women are not dangerous things: gender and categorization. *Harvard Working Papers in Linguistics* 12.

Reid, Nicholas J. 2011 [1990]. *Ngan'gityemerri: a language of the Daly River region, Northern Territory of Australia*. Lincom Europa, München.

Rijkhoff, Jan. 1991. Nominal aspect. *Journal of Semantics* 8 291–309. https://doi.org/10.1093/jos/8.4.291.

Seiler, Hansjakob. 1986. *Apprehension: language, object, and order. Part III: The universal dimension of apprehension*. Gunter Narr Verlag, Tübingen.

Steinmetz, Donald. 1986. Two principles and some rules for gender in German: inanimate nouns. *Word* 37: 189–217.

Steinmetz, Donald. 2006. Gender shifts in Germanic and Slavic: semantic motivation for neuter. *Lingua* 116: 1418–1440.

Sugiyama, Lawrence S. 2016. Physical attractiveness: an adaptationist perspective. In *The handbook of evolutionary psychology. Vol. 1: Foundations*, edited by David M. Buss, pp. 317–384. Wiley, Hoboken, NJ.

Talbot, Mary. 2019 [1998]. *Language and gender*. Polity Press, Cambridge, 3rd edn.

Tichy, Eva. 1993. Kollektiva, Genus femininum und relative Chronologie im Indogermanischen. *Historische Sprachforschung* 106 1: 1–19.

Trosterud, Trond. 2001. Genustilordning i norsk er regelstyrt. *Norsk lingvistisk tidsskrift* 19: 29–58.

Trosterud, Trond. 2006. Gender assignment in Old Norse. *Lingua* 116: 1441–1463. https://doi.org/10.1016/j.lingua.2004.06.015.

Tryon, Darrell T. 1974. *Daly Family languages, Australia*. Pacific Linguistics, Canberra.

Weber, Doris. 1999. On the function of gender. In *Gender in grammar and cognition, Vol. I: Approaches to gender*, edited by Barbara Unterbeck and Matti Rissanen, pp. 495–509. Mouton de Gruyter, Berlin.

Wiltschko, Martina. 2012. Decomposing the mass/count distinction: evidence from languages that lack it. In *Count and mass across languages*, edited by Diane Massam, pp. 146–171. Oxford University Press, Oxford.

How weak are Romanian clitic pronouns?

Ciprian-Virgil Gerstenberger
Norwegian Centre for E-health Research

Abstract

In traditional linguistics, pronouns are divided into two classes: those that can bear word stress, coined "strong", "full" or "tonal", and those that cannot, coined "weak", "clitic", or "atonal". However, in the last decades, research on this topic has shown that items generally labeled as clitics are far more complex. Somewhere between words and affixes, these hybrid linguistic entities challenge both description and modeling. As for Romanian, the debate on weak (i.e., clitic) pronouns has been dominated by the question of their categorial status: are these items clitics or affixes? In this article, I present and scrutinize different approaches that support the claim that there are differences between proclitics and enclitics, i.e., between clitics occurring before vs. after the verb; this includes not only positional, but also featural differences. I identify various types of ambiguities in Romanian that could lead to improper data interpretation, and, based on an analysis of syllabicity – the most salient feature of Romanian weak pronouns – I refute claims for treating clitics in preverbal position differently than in postverbal position. Furthermore, using evidence from both historical data and data pertaining to language varieties, I show regularities in the Romanian weak pronoun system, bringing evidence against the claim that Romanian weak pronouns show a great deal of idiosyncrasies.

Keywords: Romanian, clitics, clitic pronouns, weak pronouns, description, rule-based model

1. Introduction

Browsing through the vast literature on clitics, one can hardly find an article or a book that does not highlight the complexity of the topic and the difficulties of attempting to pin down the very nature of clitics. Somewhere between words and affixes, the label "clitic" has been applied to items that are not clearly words, nor clearly affixes (cf., for instance, Spencer and Luís 2012).

The originally simple concept of an item 'leaning' on a host either from one side, as a proclitic, or the other, as an enclitic, was not enough to describe the extensive variety of hybrid forms and their occurrences. Even the fact that clitics may occur in a clitic cluster challenges the idea of clitic-hood. In a sequence of two clitics preceding the host, it is only the second clitic which truly 'leans' on the host, while the first clitic can only 'lean' on the following clitic. The same issues with the simple clitic-host notion arise around what is referred to as mesoclitics in Portuguese (cf. Rouveret 1999), or endoclitics in Udi[1] (cf. Harris 2002). Conceptually, one and the same clitic cannot 'lean' on both sides of a host at the same time, but this is the case both with mesoclitics – which occur between verb stem and affixes – and with endoclitics – which split the root into two parts: both constructions challenge the Lexical Integrity Hypothesis (cf. Di Sciullo and Williams 1987) unless they are regarded as affixes – as Haspelmath (2022:p. 31) does with the items in the verbal complex in the Romance languages. As I will show in Section 2, for Romanian, these problems – intensified by different types of ambiguities in the orthographic system – have led to difficulties in accurately describing the clitic-host relationship.

1.1. Description issues

There seems to be a general pattern in detailed descriptions of clitics, according to which the assertion of a clitic property is accompanied by exceptions, as the following list exemplifies (cf. Caink 2006):

1. **Rigid order** Clitics, as affixes, appear in rigid order, yet there are counterexamples, e.g., Bonet (1991) for Catalan or Săvescu Ciucivara (2009) for Romanian.
2. **Stress** One of the most prominent features of clitics and the feature that coined the terms 'weak' or 'atonal' for pronouns is stress. Clitics cannot bear accent or stress, though there are plenty of counter-examples of stressed clitics in the vast literature, e.g., Klavans (1995), Ordóñez and Repetti (2006) or Săvescu Ciucivara (2009).

[1] Udi is a Nakh-Daghestanian language of Azerbaijan and Russia.

 Nordlyd 46.1: 37-57, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6341

3. **Coordination** Syntactic rules of coordination never apply for clitics, and yet, examples of coordinated clitics are found in the literature, e.g., Finocchiaro (2005) for French or Săvescu Ciucivara (2009) for Romanian.
4. **Ellipsis** Clitics are not affected by ellipsis, yet there is evidence for the opposite, e.g., Franks and King (2000) for Serbo-Croatian or Finocchiaro (2005) for Italian.

Based on these descriptions, it is not clear whether the assertion that some clitic behavior contradicts the general view on clitics (e.g., clitic coordination in Romanian) is actually a part of the standard language or it is only the idiosyncratic view of a linguist or of a linguist's informant(s). This circumstance becomes apparent for non-native speakers when discrepancies between native speakers' opinions about the grammaticality of a language sample emerge (see, for instance, Bošković 2001:p. 122, footnote 25).

Cardinaletti and Starke (1999) propose a tripartite model that distinguishes between strong, weak and clitic pronouns in terms of *structural deficiency*. They argue that Italian stressed enclitics are weak pronouns distinct from other clitics. Yet, Manzini (2014) maintains that it is "not obvious that the intermediate series ('weak') displays consistent characteristics" (for a similar view, see also Pescarini 2018). Moreover, Cardinaletti and Starke (1999)'s tripartition of pronouns, deemed by the authors as *universal*, does not fit the Romanian data, as argued for in Somesfalean (2007:p. 6). The description of Romanian data in this study brings additional support to Haspelmath's (2015) view on Cardinaletti and Starke's (1999) model of structural deficiency: "their tripartition of person forms into *clitics*, *weak pronouns* and *strong pronouns* is really based on a few interesting converging observations for German and Italian, and cannot be extended to many other languages without encountering the familiar problems" (op. cit. p. 276, footnote 2).

Regarding clitic coordination, Săvescu Ciucivara (2009:p. 24) argues that, in Romanian, "certain pre-verbal clitics can be coordinated (though not all speakers accept it)". A different perspective is offered by Dindelegan (2013:p. 388), who states that "[i]n informal registers, Romanian […] allows auxiliary and second clitic deletion in coordinated structures. […] gapping of the verb is allowed if it is repeated in the two coordinated structures." Again, there is a report on some speakers' judgements[2] on clitic coordination in Cardinaletti (1999:p. 39): "Benincà and Cinque [(1993)] report that under special prosodic and pragmatic conditions, some speakers of French and Rumanian accept coordination of two clitic pronouns in proclisis contexts, but never in enclisis contexts." Without mentioning any special prosodic or pragmatic conditions, Luraghi (2017:p. 190) claims that "proclitics can be coordinated in French and Romanian." Interestingly, while for Dobrovie-Sorin (1994:p. 61), the coordination of preverbal clitics is "marginally acceptable", for Avram (1997:p. 159), it is "not advisable".[3] In turn, Dobrovie-Sorin and Giurgea (2013:p. 262) asserts that "[c]litics cannot be coordinated, modified or focalized". It is also worth mentioning that the vast majority of Romanian grammars for native speakers do not mention this phenomenon at all. Since there are – to the best of my knowledge – no corpus- or usage-based studies on clitic coordination in Romanian, it is hard to get a clear picture.

Concerning stressed Romanian clitics, Săvescu Ciucivara (2009:p. 24) states that "[w]hen coordination does happen, both clitics are stressed", a statement that contradicts the claim that "[c]litics do not have a word accent" made by Dobrovie-Sorin and Giurgea (2013:p. 262).

As for rigid order, Săvescu Ciucivara (2009) claims that in postverbal contexts, sequences of case-syncretic plural clitic forms *ne*$_{\text{\_1.pl.acc/dat}}$ and *vă*$_{\text{\_2.pl.acc/dat}}$ – both *GERUND-ne-vă* and *GERUND-vă-ne* – are grammatically accepted, however, only with the interpretation Acc>Dat,[4] a pattern that contradicts the Dat>Acc positioning of Romanian clitics. This statement has also been mentioned by Dobrovie-Sorin and Giurgea (2013:p. 260). Yet, this allegedly grammatically correct structure is contradicted by any grammar of Romanian, and, as a matter of fact, also by Dobrovie-Sorin and Giurgea (2013) themselves in a book section titled "The order of co-occurring dative and accusative clitics": "Regardless of their position relative to the verb (pre- or post-posed), clitic pronouns form a syntactic unit inside which the order of clitics is fixed, and no insertion is allowed" (op. cit. p. 256). Since the grammaticality judgments of these examples "are based on elicitation and not on corpus research" (Dobrovie-Sorin and Giurgea 2013:p. 260),[5] a series of questions arises, questions that also apply to the claims on Romanian clitic coordination as reported in Cardinaletti (1999:p. 39). Was it an online survey or an experiment? What was the exact number of participants? What was their background and education? If it was a psycholinguistics experiment, was the presentation of examples timed or self-paced? Which distractors were used to prevent participants from giving biased answer, and how were these used? What do statistical analyses result in? Compared to Săvescu Ciucivara (2009), Nevins and Săvescu (2008) provide a clear description of

[2] as reported in Benincà and Cinque (1993)

[3] "[C]oordinarea a două forme neaccentuate […] **nu este recomandabilă**" (my emphasis).

[4] Throughout this study, I use ">" to mark the relative position in the string between two items such as Dat>Acc: the dative clitic occurs before the accusative clitic.

[5] while Săvescu Ciucivara (2009) does not describe the experimental setup of the data collection nor an evaluation of the outcome figures

the experimental setup, outcome figures, and evaluation. It is even more remarkable that, in Nevins and Săvescu (2008), there is absolutely no mention of the acceptability testing of combinations of *ne* and *vă*.

1.2. Terminology issues

In descriptions of the phenomena under scrutiny, there are other difficulties, that related to the use of terminology. For instance, van Riemsdijk (1999a:p.20) labels specific pronouns in Dutch and other Germanic languages as "hostless clitics". Yet conceptually, a clitic requires a host to *lean* on, as this is one of its defining features, and otherwise calling it a clitic does not make any sense. To take another example, synonyms for the same term can suddenly have different meanings due to re-defining them for a specific theoretical model, as is the case with the terms "weak pronoun" vs. "clitic pronoun" in Cardinaletti and Starke (1999). These two terms are used synonymously otherwise, and they are still used as synonyms in the literature that doesn't adopt Cardinaletti and Starke's (1999) view. Without any explicit indications, the reader has to discern between different contexts of use in order to understand the intended meaning of the terms on his/her own. In particular, while for Klein (2007) clitics form a proper subset of weak pronouns – namely only those items that 'phonologically clitizise' – for Cardinaletti and Starke (1999), clitics and weak pronouns are disjointed sets.

In the same vein, Cherecheş (2014) employs terms such "affixal clitics" or "internal clitics" for her prosodic model, despite a footnote in which she mentions "that this usage of the term "clitic" is purely phonological, divorced from the morphosyntactic properties of the element in question" (op. cit. p. 57, footnote 9). Nevertheless, this very specific usage of the term "clitic" obfuscates a clear distinction between syllabic and asyllabic items in the verbal complex, hence veiling a clear distinction between *supporting* and *supported* items. In particular, in the model advocated by Cherecheş (2014) "the auxiliary acts as an affixal (en)clitic when preceded by a pronominal" (op. cit. p. 57), whereas in terms of syllabicity, the vowel-initial auxiliary is always the syllabic host when preceded by a pronominal (e.g., ex. 35, and see the description of obligatory sandhi in Section 2.4.2). Conversely, Dobrovie-Sorin (1994:p. 49) promotes the term "syntactic clitics", while Legendre (2001) the term "verbal clitics" for those entities that Zwicky (1977) labeled "special clitics", e.g., clitic pronouns in Romance.

Furthermore, similar to mesoclitics in Portuguese, Romanian exhibits some structures for expressing wishes, curses, or blessings – remnants of older stages of the language – as referred to in Dobrovie-Sorin and Giurgea (2013:p. 253), Bošković (2001:p. 123, footnote 27), Dindelegan (2016), or Hill and Alboiu (2016). These structures are hardly ever described as mesoclitics in Romanian literature. The reason for this is that, in the configuration VERB>CL-PRON>VERBAL, the VERBAL item is interpreted as an inflected auxiliary verb (the form *ai* in the following Romanian example), and not as an affix of the infinitive main verb (the form *ias* in the Portuguese counter-example), as commonly found in the literature on Portuguese: Romanian *cumpăra-mi-ai cărți* vs. Portuguese *comprar-me-ias livros*[6] 'you would buy me books/may you buy me books'. In international literature however, the very same Romanian structures are inconsistently addressed, either as mesoclitics in Gerlach (2002:p. 57) or as endoclitics in Dobrovie-Sorin (1994:p. 78-79). As already mentioned, mesoclitics occur between verb stem and affixes, while endoclitics split the verb root in two parts. Linguists working with Portuguese or Udi would disagree on the synonymous use of these two terms, which highlights the confusion surrounding analyses of Romanian clitic pronouns.

1.3. Classification issues

Given this description of the various problems and contradictions concerning items labeled *clitic*, it is intriguing to try to bring order into the realm of clitic-hood, i.e., to attempt to identify – in a consistent way – sub-classes of phenomena among items referred to as clitics.

A pioneering account for systematization of the items under the umbrella term *clitics* is Zwicky (1977), who identifies three common types: (1) *special clitics* – unaccented bound forms that are variants of free forms, yet with a 'special' syntax (e.g, Romance clitic pronouns), (2) *simple clitics* – unaccented, phonologically reduced variants of full forms, occurring in the same position as the phonologically full form (e.g., the English *'ll* as in *I'll do it* as opposed to *I will do it*); and (3) *bound words* – always unaccented and phonologically subordinated to a neighboring word (e.g., English genitive *'s*).

In their influential article, Zwicky and Pullum (1983) proposed a suite of tests to distinguish between clitics and affixes (including host selectivity, arbitrary gaps, morphophonological idiosyncrasies, etc.), but Anderson (2005)

[6]example retrieved online 2022-03-10 at URL https://ciberduvidas.iscte-iul.pt/consultorio/perguntas/pronomes-mesocliticos/7865

critically states that "these points are merely descriptive observations about differences in the behavior of two pre-systematically understood classes" (op. cit. p. 33). Despite the fact that these criteria present only tendencies useful for defining what is or is not a clitic, many linguists have used them as solid arguments for discriminating between clitics and affixes (for Romanian, e.g., Monachesi 2001). Even Zwicky himself highlights the distinctions between diagnostic tests and defining criteria in Zwicky (1985:p. 285): "what is normally intended, when such tests are appealed to, is more analogous to medical diagnosis than to operations using an axiomatic system." An overview of different taxonomic efforts towards a general characterization of the category of clitics is given in Haspelmath (2015:p. 275). At the same time, Haspelmath's article is a useful point of reference for an in-depth consideration of the use of grammatical terms cross-linguistically by means of clitic-hood.

In terms of syntactic categories, clitics can belong to various types, such as pronouns, determiners, auxiliaries, negation markers and interrogative markers (cf., e.g., Zwicky 1977). With the complexities concerning clitics and clitic-hood in mind, as presented here and in the previous sections, the remainder of the present study focuses on a single part-of-speech category in a specific language, namely Romanian pronominal clitics, which are weak pronouns in the context they occur in (i.e., the verbal complex).

2. Romanian Weak Pronouns

2.1. Previous approaches

For a long time, the focus of research on Romanian clitics has been on the pronominal items with various labels such as "atonal", "clitic", or "weak", depending on the source. The linguistics literature on Romanian weak pronouns (hereafter RWPs) features a great diversity of descriptions and models. One of the first in-depth corpus-based descriptions of the Romanian verbal complex is provided by Bredemeier (1976), who also works out a detailed and accurate theory-neutral formalization in terms of context-derived constraints. Avram (1986) offers a broad depiction of sandhi in RWPs. Barbu (1999) provides a description of the verbal complex, Dobrovie-Sorin (1999a) a generative approach to the syntax of RWPs, and Somesfalean (2007) an approach to argumental pronominal forms based on data from Romanian and other Romance languages couched in the theoretical framework of the Minimalist Theory. Săvescu Ciucivara (2009) offers another generative perspective on the syntactic analysis of pronominal clitic combinations and ordering in Romance, especially in Romanian. Calude (2001) compares Romanian to French and Serbo-Croatian clitics and concludes that Romanian clitics share many more features with their Serbo-Croatian than with their French counterparts. Furthermore, Monachesi (2001) and Monachesi (2005) deal with RWPs using Head-Driven Phrase Structure Grammar, Barbu and Toivonen (2018) models direct object RWPs within a Lexical Functional Grammar framework, while Klein (2007) treats them within the Dynamic Syntax formalism. Optimality Theory is represented by a series of models such as Popescu (2000), Sasaki and Căluianu (2000), Legendre (2001), Popescu (2003), and Chereches¸ (2014).

There is a long-standing debate about the categorial status of RWPs, specifically concerning whether they are clitics or affixes. While Barbu (1999) and Monachesi (2001) put forward arguments for classifying RWPs as affixes, Popescu (2003) and Gerlach (2002) label them as clitics, irrespective of the RWPs' position towards the verb, i.e., both proclitics and enclitics have the same categorical status as either affixes or clitics. However, some scholars such as Benincà and Cinque (1993) or Cardinaletti and Starke (1999) have promoted the view that there is a fundamental difference between clitic pronouns occurring pre-verbally and those occurring post-verbally; in other words, positing not only a difference in items' position relative to the verb, but also a difference in their other properties.

The threefold aim of the following exploration of the syntax and phonology of RWPs is to present evidence that: (1) there is no basic feature difference between pre- and post-verbal RWPs; (2) it is possible to do away with alleged idiosyncrasies; (3) a thorough, careful description of the data provides a solid foundation for the implementation of a computational linguistics model for generating RWP surface forms.

Note that Appendix 1 and Appedix 2 present an extensive list of examples to provide an overview of the data and facilitate comparisons of individual instances. Unless otherwise specified, example numbers throughout the paper refer to the examples in these appendices.

2.2. Two levels of clitic-hood

The Romanian pronominal system is best described as a tripartite model: strong pronouns (not subject of this study) vs. weak (or clitic) pronouns. The weak pronouns surface in two different forms: syllabic vs. asyllabic (see Table 1).

Due to the ambiguities described below in Section 2.5, the difference between a syllabic and an asyllabic RWP is not always manifest. This issue leads to an inadequate classification of RWPs in terms of syllabicity, which in turn, leads to difficulties in clearly differentiating between host and clitic at the morphophonological level. Moreover, occasionally in the literature on RWPs, it is not clear whether it is about a host at the morphophonological level or at the phrasal level: is the host-clitic pair described in terms of syllabicity or in terms of phrasal stress?

Both Dobrovie-Sorin (1999b) and Klein (2007) point out the need to differentiate between two levels of RWP descriptions that are independent of each other: a phonological/morphophonological level and a syntactic level. The two-level view on ciltic-hood for Romanian presented in this study is congruent with their view.

At the phrasal level, all weak pronouns and other non-stressable – mostly monosyllabic – items in the verbal complex are clitics to the verb,[7] i.e., the prosodic phrase host. This means that, at the morpho-phonological level, an asyllabic RWP form can be a 'clitic' to a neighboring syllabic RWP – the 'syllabic host' – and at the same time, both are proclitics or enclitics to the verb at the phrasal level.

As for the syllabic level, instead of using the terms "phonological" or "morpho-phonological" cliticization, I use the general term "sandhi" and explicitly mark the syllabic support as "syllabic host" and the asyllabic item as "asyllabic clitic" to avoid possible ambiguities.

2.3. Syntactic features of Romanian Weak Pronouns

As with clitic pronouns in other Romance languages, RWPs occur in the verbal complex before the verb as proclitics or after the verb as enclitics. They can form a cluster of up to three RWPs with a fixed order within the clitic cluster: Dat>Dat>Acc, independently of the relative position of the RWP cluster towards the verb (see Dobrovie-Sorin and Giurgea 2013:p. 257ff or Jianu 2013).

In preverbal position, only auxiliaries and a small set of monosyllabic adverbs can intervene between RWPs and the main verb, while in postverbal position, RWPs immediately follow the main verb. In modern Romanian, the auxiliary occurs mostly in preverbal position, however, in wishes, curses, or blessings as well as in vernacular language, the auxiliary may also occur in postverbal position (cf. Gerlach 2002:p. 57). The order of the RWPs cluster and the auxiliary is the same both preverbally (ex. 36) and postverbally (ex. 37), namely RWPs>AUX.

Figure 1 shows an instance of a verbal complex with preverbal RWPs in a maximal context of other possible occurring items, such as negation, auxiliary, monosyllabic adverb intensifiers. Adapted from Cherecheş (2014:p. 51) and Dobrovie-Sorin and Giurgea (2013:p. 257), the example shows the fixed order of the items within the verbal complex.[8]

[Neg > Dat-RWP > Dat-RWP > Acc-RWP > Tense-/Mood-Aux > Adv > Perfective > Adv > Verb]

[Nu	mi	ţi	l-	ar	mai	fi	tot	aruncat]	vrăjitoarea	încoace	şi -ncolo.
[not	me$_{dat}$	you$_{dat}$	him$_{acc}$	have$_{cond.3.sg}$	anymore	be$_{perf}$	continually	thrown]	the witch	here	and away.

'The witch [wouldn't have kept throwing him] back and forth.'

Figure 1: Preverbal RWPs in [a maximal verbal complex]

Clitic placement RWPs occur in preverbal position with finite (examples 3, 12, 15), infinite (examples 19, 20), and negated imperative verb forms (examples 23, 22), while in postverbal position with participle/gerund (ex. 21) as well as with non-negated imperative verb forms (examples 18, 25, 27).[9] Depending on the formulation, affectionate exclamations, wishes, blessings, and curses can occur in preverbal position (ex. 36) or postverbal position (ex. 37). Due to its phonological shape as /o/, the RWP for 3.sg.acc.f exhibits unique behavior. Preverbally, it occurs only if there is no auxiliary starting with a vowel (ex. 34), otherwise, it occurs postverbally (ex. 35).

[7] In the case of Romanian weak copula verb forms, the prosodic phrase host is the predicative (cf. Section 2.4.2).

[8] The reason why the two dative clitics do not appear in the English translation is because they both can be interpreted as dativus ethicus, which is difficult to express in English. In Romanian, constructions with dativus ethicus are not employed too often nowadays; for their interpretation, see Dobrovie-Sorin and Giurgea (2013:p. 257ff) and Jianu (2013:p. 257ff).

[9] for a comparison to other Romance languages, see Gerlach (2002:p. 267)

Clitic doubling and differential object marking As Spanish, Romanian features clitic doubling, i.e., when an object surfaces twice in specific constructions, both as a clitic and as a full pronoun or noun. In such configurations, the accusative object is marked with the preposition *pe* in Romanian, the counterpart to the Spanish marker *a*. This phenomenon is called *Differential Object Marking* (cf. Tigău 2021).

Combinatorial restrictions As with other Romance languages, RWPs feature arbitrary gaps in clitic-clitic combinations, a phenomenon coined *Person Case Constraint* (cf. Bonet 1994).

Clitic climbing In a few cases, as with the modal *a putea* 'to be able to', Romanian exhibits clitic climbing (ex. 31), where the clitic occurs before the modal instead of the main verb (cf. Monachesi 2005:p. 206).

Coordination As an argument for the difference between proclitics and enclitics, Benincà and Cinque (1993:p. 2323-2324) purport that in Romanian proclitic pronouns can be coordinated, which is not possible with enclitics. However, as mentioned above, RWPs cannot be coordinated, modified or focalized (cf. Dobrovie-Sorin and Giurgea 2013:p. 262), but see the debate on that specific topic presented in Section 1.1. If two coordinated verbs share the same clitic, the clitic has to be repeated for each verb, i.e., the scope of pronominal clitics does not extend across coordination (cf. Monachesi 2001:p. 264).

Interpolation A further syntactic argument for the difference between proclitics and enclitics put forward by Benincà and Cinque (1993:p. 2324-2325) is the interpolation, the occurrence of adverbs or other constituents between a clitic and its host, which is possible in proclitic, but not in enclitic contexts. Yet, to conclude that there is a difference between proclitics and enclitics because of interpolation phenomena is a fallacy: from the fact that the interpolated adverbs have a fixed position, namely right before the main verb, does not follow that there is a difference between proclitic and enclitic pronouns. Given the fix position of the clitic items Dat-RWP>Acc-RWP>AUX both before and after the verbal host and following the same line of reasoning as Benincà and Cinque (1993), one could claim that there is a closer relation between the auxiliary and the main verb in preverbal position (ex. 36) than in postverbal position (ex. 37), where the Acc-RWP occurs between the main and the auxiliary verb. Or that there is a closer relation between an Acc-RWP and the main verb in preverbal position (ex. 26) than in postverbal position (ex. 27), where the Dat-RWP occurs between the main verb and the Acc-RWP. Besides, Pescarini (2019) points to empirical data featuring interpolation phenomena also in enclitic contexts.

2.4. Phonological features of Romanian Weak Pronouns

In order to identify phonological constraints on RWP combinations, one should first categorize the phonological shapes of RWPs based on the different contexts they occur in (for more details, see Avram 1986 or Popescu 2000). Although there is no disagreement about partitioning RWPs into two categories (one for syllabic and one for asyllabic forms), there is a great variety in terminology. Some linguistic RWP descriptions offer a dichotomy between *free* vs. *bound* – as in Iliescu (1975:p. 51), Guţu-Romalo (2008:p. 203), and Dindelegan (2013:p. 382) – between *full* vs. *short* – as in Nastasenco (1997:p. 19) – between *full* vs. *reduced* – as in Popescu (2000:p. 775ff), and Mišeska Tomić (2006:p. 280) – between *full* vs. *non-full* – as in Calude (2001:p. 98) – between *long* vs. *short* – as in Cherecheş (2014:p. 52) – or between *syllabic* vs. *asyllabic* – as in Dobrovie-Sorin and Giurgea (2013:p. 261). Since RWPs can be either syllabic or asyllabic, and these terms adequately and succinctly capture on aspect of the categorical essence of RWPs, I use the terms "*syllabic*" vs. "*asyllabic*".

2.4.1. Î-Prothetic forms

Prothesis, the addition of a sound at the beginning of a word without changing its meaning, is not unusual in Romance languages. In Romanian, *î*-prothesis[10] is encountered with a series of RWPs that have to surface as asyllabic forms and thus need syllabic support, such as *îmi* [ɨm^{i}] or *îți* [ɨts^{i}] (cf. Lombard 1976). In Romanian grammar, these forms are always described as syllabic.

[10]The emergence of Romanian *î*-prothesis occurred in a period between the thirteenth and the sixteenth century (cf. Dindelegan 2016:p. 67).

Notwithstanding the orthography of *î*-prothetic forms, I argue that a more appropriate RWP description abstracts away from the prothetic *î*, classifying these forms as asyllabic. There are three arguments that support this claim.

First, the syllable structure of the *î*-prothetic RWP forms are the same as in other combinations of <syllabic support>-<asyllabic form>, independent of the type of the supporting item. The syllables *îmi* [ɨm^{i}], *și-mi* [ʃim^{i}], *nu-mi* [num^{i}], *că-mi* [cəm^{i}], *să-mi* [səm^{i}], and *dă-mi* [dəm^{i}] have the same structure – <non-pronominal syllabic support> followed by <pronominal asyllabic form> – despite the fact that each syllabic support item is of a different kind: *î* is a prothetic vowel, *și* is the conjunction 'and', *nu* is a negation particle, *că* is a subordinator, *să* is a subjunctive marker, and *dă* is the form |give$_{\text{2p.sg.imp}}$| of the verb *a da* 'to give'. Yet, the asyllabic weak pronoun part, i.e., the string carrying the semantics of the pronoun, is the same in all these combinations, namely *mi* [m^{i}] |cl$_{\text{1.sg.dat}}$|.

Second, the prothetic *î* is required only in very specific contexts: when each of the asyllabic forms *mi* [m^{i}], *ți* [ts^{i}], *i* [i̯], *și* [ʃ^{i}], or *l* [l] occurs alone and there is no vocalic support to the left or to the right. In contexts allowing optionality, the *î*-prothetic forms 'compete' with appropriate syllabic support items (ex. 22 vs. 23 or ex. 42 vs. 43), as described in Section 2.4.3.

Third, *î* is a prothetic vowel not only for weak pronouns but also for weak forms of the verb *a fi* 'to be', such as *îs* [ɨs] in *îs pe munte* 'I am/they are on the mountain' or *îi* [ɨi̯] in *îi acasă* 'he/she is home', as mentioned in Lombard (1976:p. 116), Rosetti (1986:p. 373), or Avram (1986:p. 558). These weak *î*-prothetic verb forms – nowadays used in regional varieties and colloquial language (cf. Avram 1986:p. 558 or Zafiu 2019) – have exactly the same occurrence constraints as the *î*-prothetic RWPs.

These facts demonstrate that treating the prothetic *î* as it is, namely, just prothesis, simplifies the categorization of RWPs by reducing their inventory. Moreover, the abstraction from the prothetic *î* does without Dobrovie-Sorin and Giurgea's (2013:p. 261) **Inside-** vs. **Outside clitic clusters** for dative-syllabic, which is a mixture of form and context. The optimized surface form description of RWPs is presented in Table 1.[11]

2.4.2. Obligatory sandhi

Having clarified the status of *î*-prothetic forms as asyllabic, I now turn my attention to the phonological constraints on RWP combinations to distinguish possible regularities, i.e., patterns, in clitic clusters. As mentioned above, I use the term "sandhi" to refer to phonological cliticization, i.e., the combination of an asyllabic RWP and a syllabic host.

At the beginning of this article, I noted that I use evidence from both historical data and data pertaining to language varieties. As mentioned above in Section 2.4.1, this is actually the case with the Romanian weak verb forms (hereafter RWVs). These forms are *s* [s] – `1.sg/pl.pres` of the verb *a fi* 'to be' with the strong form *sunt*[12] – and *i* [i̯] – `3.sg.pres` with the strong form *e/este*.[13] RWVs can combine with dative RWP forms and can appear both before – *mi-s acolo* in ex. 32 – and after – *acolo mi-s* in ex. 33 – the stressed word *acolo* [a.ˈco.lo] 'there' within the verbal construction (cf. Iliescu 1975:p. 59).

Ignoring any context that enables optional sandhi, which is treated separately in Section 2.4.3, there are two configurations relevant for the description of obligatory sandhi: (1) contexts with no monosyllabic vowel-initial item immediately following the RWP cluster; (2) contexts with an auxiliary starting with a vowel or with the *o*-RWP right after the RWP cluster.[14]

To find out whether there are differences in syllabicity between identical clitic clusters in pre- vs. post-verbal contexts, each RWP combination has to be taken into account. In the case of RWP-RWV combinations, it is about pre- vs. post-predicative contexts. Due to the parameters `RIGID ORDER` and `MAXIMAL NUMBER`, the following possible configurations have to be examined: a single Acc-RWP, a single Dat-RWP, the sequence Dat-RWP>Acc-RWP, the sequence Dat-RWP>Dat-RWP, and the sequence Dat-RWP>Dat-RWP>Acc-RWP. Moreover, due to the similar behavior between RWVs and *î*-prothetic RWPs, contexts with a single RWV, the sequence Dat-RWP>RWV, and the sequence Dat-RWP>Dat-RWP>RWV are relevant, too.

[11] for similar synopses, see (Avram 1986:p. 554) or Dobrovie-Sorin and Giurgea (2013:p. 261)

[12] spelled *sînt* before the orthographic rules adopted in 1993 by the Romanian Academy

[13] There are other weak verb forms – such as the future auxiliary for `2p.sg` *îi* and for `2p.pl` *îți* (cf. Avram 1986:p. 558) – which can be modeled on a par with the present weak verb forms treated here. These future auxiliary forms are homonymous with the RWPs *îi* and *îți*, respectively.

[14] Since the *o*-RWP behaves like a vowel-initial auxiliary and is treated accordingly, references to an RWP cluster in this study do not include this item.

Case	*Number*	*Person*	*Type*	*Gender*	*Syllabic*	*Asyllabic*	
						onset	***coda***
A	*Sg*	*1.*	*pers/refl*	*m/f*	mă [mə]	m [m]	—
c		*2.*	*pers/refl*	*m/f*	te [te]	te [te̯]	—
c		*3.*	*pers*	*m*	—	l [l]	l [l]
u				*f*	o [o]	o [o̯]	—
s			*relf*	*m/f*	se [se]	se [se̯], s [s]	—
a	*Pl*	*1.*	*pers/refl*	*m/f*	ne [ne]	ne [ne̯]	—
t		*2.*	*pers/refl*	*m/f*	vă [və]	v [v]	—
i		*3.*	*pers*	*m*	—	i [i̯]	i [i̯]
v				*f*	le [le]	le [le̯]	—
e			*relf*	*m/f*	se [se]	se [se̯], s [s]	—
	Sg	*1.*	*pers/refl*	*m/f*	mi [mi]	mi [mi̯]	mi [m^{i}]
D		*2.*	*pers/refl*	*m/f*	ți [tsi]	ți [tsi̯]	ți [ts^{i}]
a		*3.*	*pers*	*m/f*	i [i]	i [i̯]	i [i̯]
t			*relf*	*m/f*	și [ʃi]	și [ʃi̯]	și [ʃ^{i}]
i	*Pl*	*1.*	*pers/refl*	*m/f*	ni [ni], ne [ne]	ni [ni̯], ne [ne̯]	—
v		*2.*	*pers/refl*	*m/f*	vi [vi], vă [və]	vi [vi̯], v [v]	—
e		*3.*	*pers*	*m/f*	li [li], le [le]	li [li̯], le [le̯]	—
			relf	*m/f*	și [ʃi]	și [ʃi̯]	și [ʃ^{i}]

Table 1: **Surface forms of Romanian weak pronouns**: orthographic form [IPA]

I. Contexts without a monosyllabic vowel-initial item following the RWP cluster

A. RWPs/RWVs in the right-most position

A1. consonants: Always asyllabic, *l*-RWP and *s*-RWV occur only in the right-most position.

A2. *i*-forms: All dative singular *i*-RWPs, the accusative plural *i*-RWP as well as the *i*-RWV occur in the right-most position as asyllabic forms, either as palatalized consonants[15] (ex. 40) or as glides (ex. 42).

A3. *e*- and *ă*-forms: All RWP forms ending in *ă* [ə] or *e* [e] occur in the right-most position as syllabic forms (ex. 12). The plural dative RWPs surface as dative-accusative case-syncretic forms,[16] as in *vă dă mere* [və.ˈdə.ˈme.re] |cl$_{\text{2.pl.dat}}$ give$_{\text{3.sg.pres}}$ apple$_{\text{acc.pl.indf}}$| 'he/she gives you apples'.

[15] In her prosodic analysis of RWPs, Cherecheş (2014) observes a parallel between RWP clusters with *i*-forms and, e.g., plural masculine nouns ending in *i*: "the plural marker for masculine nouns -i reduces to a palatalization gesture word-finally but stays a full vowel when followed by extra inflectional material" (op. cit. p. 55), for instance, *lup* [ˈlup] 'wolf' vs. *lupi* [ˈlup^{i}] 'wolves' vs. *lupilor* [ˈlu.pi.lor] 'of/to the wolves'. Avram (1986:p. 552) refers to this palatalization gesture as 'final pseudo-i'. In the context of RWPs, the right-most position in a cluster is equal to word finality in Cherecheş' (2014) example.

[16] According to Giurgea (2013), the RWP forms for plural dative such as *ni*, *vi*, and *li* emerged by a kind of dissimilation process, in analogy to the singular dative *i*-forms such as *mi* or *ți*.

B. RWPs in non-right-most positions

In all other positions, only syllabic *i*-vocalic dative RWP forms can occur (ex. 26; cf. footnote 15).

C. Clitic-host relation

C1. single asyllabic forms: For RWP/RWV items that surface as single asyllabic forms, the use of *î*-prothesis as syllabic support is obligatory in pre-verbal/pre-predicative position. In post-verbal position, the verb is the syllabic support for the asyllabic RWPs. In post-predicative position, both the prothetic *î* and the predicate can be syllabic host to an RWV.

C2. syllabic forms: For RWPs that surface as syllabic forms – alone or in a cluster – there is no need for syllabic support: both pre- and post-verbally, they occur as independent, unstressed, syllabic forms.

C3. clusters: In other contexts, i.e., in RWP/RWP-RWV clusters with two or three items, the asyllabic right-most item uses its neighbor to the left as syllabic support.

Overall picture: Any right-most obligatory asyllabic item – alone or in an RWP/RWP-RWV cluster – has its *syllabic host to the left*. The host can be an *î*-prothetic vowel (ex. 3), a syllabic RWP (ex. 24 or ex. 33), or a verb (ex. 18). Hence, in this context, all asyllabic items are enclitics at the syllabic level, independent of whether they are proclitics or enclitics at the phrasal level.

II. Contexts with monosyllabic vowel-initial items following the RWP cluster

In contexts with monosyllabic vowel-initial items in the verbal complex – such as the *o*-RWP or the auxiliary *ar* – RWV items do not occur (an accusative object is incompatible with a construction of the verb *a fi* 'to be'). As already mentioned in the syntax sketch, the *o*-RWP is an exception. If both the *o*-RWP and a vowel-initial monosyllabic auxiliary have to be expressed in the verbal complex, the *o*-RWP always occurs postverbally as a syllable (ex. 34 vs. 35; see the phonological constraint `NO HIATUS` in Popescu 2000 or Chereches¸ 2014). In seldom contexts, when both the *o*-RWP and a vowel-initial monosyllabic auxiliary occur postverbally, the *o*-RWP forms a diphthong to the following vowel as a glide, thus becoming an asyllabic proclitic just like any other RWP, as in ex. 38 (cf. also Bošković 2001:p. 123, footnote 27).

A. RWPs in the right-most position

A1. consonants: In addition to the *l*-RWP, the *ă*-forms *mă* and *vă* as well as the *se*-RWP surface in this context as [m], [v], and [s], respectively (*m-ai văzut* [ma^{i}.və.^{l}zut] |cl$_{_1.sg.acc}$ have$_{_2.sg.pres}$ see$_{_part.perf}$| 'you have seen me').

A2. *e*-forms: With the exception of the reflexive *se* (see above), all accusative and dative forms ending in *e* are obligatorily asyllabic, featuring the glide [e̯] (ex. 36).

A3. *i*-forms: All *i*-RWP forms are obligatorily asyllabic, featuring the glide [i̯] (ex. 34).

B. RWPs in non-right-most positions

In all other positions, there is no change to the previous context type, and only syllabic *i*-RWP forms in dative can occur (cf. footnote 15).

C. Clitic-host relation

All items in the right-most position surface without exception as asyllabic items, thus needing syllabic support, which, in this context, is provided by the following item. The glides form a diphthong with the following vowel, while the consonants are onsets of the syllabic host.

Overall picture: Any right-most item – alone or in an RWP cluster – is an obligatorily asyllabic item and has its *syllabic host to the right*. The host can be the *o*-RWP (ex. 34) or any monosyllabic vowel-initial item (examples 35, 36, 37, 39). Hence, in this context, all asyllabic items are proclitics at the syllabic level, independent of their position towards the verb.

This RWP analysis of obligatory sandhi indicates that (1) the RWP system is not idiosyncratic, and (2) there is no difference in syllabicity between RWP clusters in pre- vs. post-verbal position. Thus, the claim that clitics occupying postverbal position "show obligatory phonological cliticization" (Dobrovie-Sorin 1999a:p. 533) is untenable. The same applies to the claim that "[p]ostverbal weak pronouns always encliticise" (Klein 2007:p. 62), where "the process of cliticisation is a phonological process" (Klein 2007:p. 61). Neither Dobrovie-Sorin (1999a) nor Klein (2007) gives an accurate description of the differences in phonology between RWPs in pre- vs. post-verbal positions, e.g., by showing which is the *phonological host* and which the *phonological clitic* in the two environments. In particular, in ex. 26 and ex. 27, there is no phonological cliticization involved whatsoever, neither preverbally in ex. 26 nor postverbally in ex. 27: on both sides of the verb, all RWPs are syllabic. This is also illustrated by the correct analysis of RWP syllabicity for exampes 12, 13, 15 and 16 as well as the analysis of the hyphen used only as a postverbality marker in the "Disambiguation" column for ex. 13 and 16 in Appendix 1. It is obvious that both Dobrovie-Sorin (1999a) and Klein (2007) overlook a very subtle ambiguity in Romanian orthography, namely, the *hyphen ambiguity* (see Section 2.5).

As a matter of fact, it is not the clitic sequences that differ in pre- vs. post-verbal position, but some verb forms that have to adjust for specific enclitic configurations. The Romanian gerund without enclitics ends in *-nd* as in *dând mere* [ˈdɨnd.ˈme.re] |give$_{_ger}$ apple$_{_acc.pl.indf}$| 'giving apples'. This ending doesn't change when followed by the *o*-RWP: *dând-o* [ˈdɨn.do] |give$_{_ger}$ cl$_{_3.sg.acc.f}$| 'giving it/her'. However, when followed by other clitics or clitic sequences, the gerund form features a final *-u* as in ex. 21 (cf. Maiden et al. 2021:p. 149). The same is the case with verb forms ending in an asyllabic *-i* (cf. Footnote 15), which becomes syllabic when followed by enclitics: *Le dați afară.* [le.ˈdatsʲ.a.ˈfa.rə] |cl$_{_3.pl.acc.f}$ throw$_{_2.pl.pres}$ out| 'You throw them out.' vs. *Dați-le afară!* [ˈda.tsi.le.a.ˈfa.rə] |throw$_{_2.pl.imp}$ cl$_{_3.pl.acc.f}$ out| 'Throw them out!'. And again, the *o*-RWP is an exception thereof: *Dați-o afară!* [ˈda.tsi̯o.a.ˈfa.rə] |throw$_{_2.pl.imp}$ cl$_{_3.sg.acc.f}$ out| 'Throw it/her out!'. Accordingly, *u*-epenthesis phenomena in the given contexts are instances of obligatory sandhi of gerund forms.

2.4.3. Optional sandhi

When the context is favorable, optional sandhi in RWPs may emerge in contexts complementary to the obligatory sandhi ones.

A. RWPs in preverbal position

A1.: when the left-most item before a single asyllabic RWP ends with a vowel (ex. 40 vs. 41);

A2.: when the verb begins with an unstressed vowel right after a single asyllabic RWP (ex. 42 vs. 43);

A3.: when the verb begins with an unstressed vowel immediately after the right-most syllabic RWP – *le aduci* [le.a.ˈdutʃʲ] vs. *le-aduci* [le̯a.ˈdutʃʲ] |cl$_{_3.pl.acc.f}$ bring$_{_2.sg.pres}$| 'you bring them' or *mi le aduci* [mi.le.a.ˈdutʃʲ] vs. *mi le-aduci* [mi.le̯a.ˈdutʃʲ] |cl$_{_1.sg.dat}$ cl$_{_3.pl.acc.f}$ bring$_{_2.sg.pres}$| 'you bring them to me'.

Both **A1.** and **A2.** and the combination thereof – such as *că îmi aduci mere* [cə.ɨmʲ.a.ˈdutʃʲ.ˈme.re] |that î$_{_host}$cl$_{_1.sg.dat}$ bring$_{_2.sg.pres}$ apple$_{_acc.pl.indf}$| 'that you bring me apples' – illustrate the 'competition' between *î*-prothetic vowels and other contextually appropriate syllabic support items, as mentioned in Section 2.4.1. Strictly speaking, these are instances of obligatory asyllabic clitics with optional choice of syllabic host.[17] By contrast, **A3.** illustrates genuine optional sandhi, i.e, contexts where a syllabic RWP form 'competes' with its asyllabic counterpart: [le] vs. [le̯].

In the construction Dat-RWP>TO-BE, the plural dative *i*-forms combine with the verb as *i*-glides in optional sandhi, i.e., on a par with the singular dative *i*-RWP forms, as in *li-e sete* [li̯e.ˈse.te] vs. *le e sete* [le.i̯e.ˈse.te] |cl$_{_3.pl.dat}$ be$_{_3.sg.pres}$ thirst| 'they are thirsty'.

[17] Mutatis mutandis, this is the case with RWV forms in similar contexts, too.

B. RWPs in postverbal position

The only configuration that allows RWP optional sandhi in postverbal position is when a word begins with an unstressed vowel directly after the the right-most syllabic RWP (ex. 27 vs. 28).

C. Clitic-host relation

C1. [asyllabic RWP = clitic]

A single asyllabic RWP can have a host both to its left, e.g., a subordinator (ex. 40) or the prothetic *î* (ex. 41) and to its right, e.g., the main verb (ex. 42).

A right-most syllabic RWP can combine in optional sandhi only to its right: preverbally with the verb – *le-aduci* [lḛa.ˈdutʃⁱ] |cl$_{\text{_3.pl.acc.f}}$ bring$_{\text{_2.sg.pres}}$| 'You bring them.' – or postverbally with an item with non-stressed initial vowel that immediately follows the VERB-RWPs sequence – *adu-le-acolo* [a.ˈdu.lḛa.ˈco.lo] |bring$_{\text{_2.sg.imp}}$ cl$_{\text{_3.pl.acc.f}}$ there| 'Bring them there!'.

C2. [syllabic RWP = host]

Due to its sonority, the vowel *î* [ɨ] at the beginning of a word – other than RWPs with *î*-prothesis – is deleted in optional sandhi. This allows both for optional sandhi on each side of a monosyllabic vowel-only item such as **o**-RWP – *s-o-ncep* [son.ˈtʃep] vs. *s-o încep* [so.ɨn.ˈtʃep] vs. *să o încep* [sə.o.ɨn.ˈtʃep] |that cl$_{\text{_3.sg.acc.f}}$ start$_{\text{_1.sg.pres}}$| 'that I start it' – and for a co-occurrence of obligatory and optional sandhi in the same context – *le-o-ntind* [lḛon.ˈtind] vs. *le-o întind* [lḛo.ɨn.ˈtind] |cl$_{\text{_3.pl.dat}}$ cl$_{\text{_3.sg.acc.f}}$ stretch$_{\text{_1.sg.pres}}$| 'I stretch it for them'. Interestingly, in such contexts, even a single, otherwise obligatory asyllabic RWP such as *mi* [mⁱ] can surface as syllabic form *mi* [mi], replacing the initial central vowel *î* [ɨ] of the verb – ex. 44[18] vs. 45.

Such an instance of **interlocked cliticization** between the phrasal and the syllabic level is detailed in ex. 29 (vs. ex. 30) – *se-ntâmplă* 'it happens'. Here, the monosyllabic reflexive RWP *se* is the syllabic host for the asyllabic segment [n] of the verb *întâmplă*, while, at the same time, the verb itself is the phrasal host for the monosyllabic, non-stressed RWP *se*.

Stressed vowel-initial items do not allow for optional sandhi, neither preverbally – **le-aflu* vs. *le aflu* [le.ˈa.flu] |cl$_{\text{_3.pl.acc}}$ find$_{\text{_1.sg.pres}}$| 'I find them' – nor postverbally – **Dă-le-altuia!* vs. *Dă-le altuia!* [ˈdə.le.ˈal.tu.i̯a] |give$_{\text{_2.sg.imp}}$ cl$_{\text{_3.pl.acc.f}}$ other$_{\text{_sg.dat.m}}$| 'Give them to another!'. This seems to be related to the fact that RWPs cannot be stressed. However, there are some contexts of optional sandhi where the stressed negation particle *nu* loses the vowel, while the syllabic host – the immediately following monosyllabic vowel-initial item – acquires the stress, as in *n-o văd* [ˈno.ˈvəd] vs. *nu o văd* [ˈnu.o.ˈvəd] |NEG cl$_{\text{_3.sg.acc.f}}$ see$_{\text{_1.sg.pres}}$| 'I don't see her/it' or *n-am văzut-o* [ˈnam.və.ˈzu.to] vs. *nu am văzut-o* [ˈnu.am.və.ˈzu.to] |NEG have$_{\text{_1.sg.pres}}$ see$_{\text{_part.perf}}$ cl$_{\text{_3.sg.acc.f}}$| 'I haven't seen her/it'.

Note that optional sandhi between shorter – usually monosyllabic – items is much more prevalent than between a monosyllabic and a heavy polysyllabic item (cf. also Gerlach 2002:p. 141). That means that, if there is a choice for optional sandhi between two items with different weights, the combination to the shorter one might be preferred.[19]

[18] Original as *roua dimineții mi-mbată inima* 'the morning dew makes my heart drunk' retrieved from the URL https://poeziipentrusufletulmeu.com/2019/09/07/dorule/ on 2022-02-27, but, since *dimineții* is not an essential part of the example, I have left it out for reasons of space.

[19] Although further discussion of this topic would go beyond the scope of the current study, an intriguing question concerns the contexts in which optional sandhi is chosen. When does a speaker decide on a variant with optional sandhi and when not? Is optional sandhi truly optional, or are there some conditions involved that we have yet to recognize? There are contradicting opinions on this topic, e.g., Popescu (2003), claiming that the trigger for optional sandhi is **speech rate** vs. Dindelegan (2013:p. 388), claiming that optional sandhi is controlled by **language register rules**.

2.5. Orthography issues

Writing is the most prominent medium of communication in science. Hence, writing and notation systems, both for natural language and for specific scientific domains, are crucial for a correct understanding of messages, ideas, and argumentation. Yet these systems are not perfect, and they are prone to changes and improvements. The orthography of a language can be designed with a phonetic or an etymological principle in mind, and can be easier or more difficult to master, even by native speakers (cf. Firică 2009). The Romanian orthographic system is no exception.

2.5.1. Ambiguities

Romanian orthography exhibits various types of ambiguity, which hinder an easy understanding of the RWP data, such as different types of homonymy as well as hyphen ambiguity.

For instance, there is homonymy[20] – actually, both homophony and homography – between the indefinite article `fem.sg.nom-acc` *o*, the cardinal numeral feminine *o*, the future particle *o*,[21] and the RWP *o*, as in *o comisie o să o vadă numai o zi* [o.co.ˈmi.si.e.o.sə.o.ˈva.də.ˈnu.maɨ̯.o.ˈzi] |o$_{\text{art.sg.indf.f}}$ committee o$_{\text{part.fut}}$ that cl$_{\text{3.sg.acc.f}}$ see$_{\text{3.sg.conj.pres}}$ only o$_{\text{card.num.f}}$ day| ‘a committee will see her only one day’. Another instance of the same type is the dative-reflexive RWP *și* and the conjunction *și* ‘and’, as in *și le cumpără* ***și*** *și le revinde* [ʃi.le.ˈcum.pə.rə.ʃi.ʃi.le.re.ˈvin.de] |cl$_{\text{3.sg.dat.refl}}$ cl$_{\text{3.pl.acc.f}}$ buy$_{\text{3.sg.pres}}$ and$_{\text{conj}}$ cl$_{\text{3.sg.dat.refl}}$ cl$_{\text{3.pl.acc.f}}$ resell$_{\text{3.sg.pres}}$| ‘he/she buys them for him-/herself **and** resells them for him-/herself’.

Grapheme-phoneme ambiguity is evidenced, e.g., by the orthographic form *mi*-RWP for `1p.sg.dat`, which can be either the syllabic form [mi] (ex. 24), the asyllabic form with a glide [mɨ̯] (ex. 34), or the asyllabic palatalized form [mʲ] (ex. 40).

There is a particularly treacherous homonymy between the *ți*-RWP and the imperative plural suffix *ți* as in *Pune-ți-l jos!* [ˈpu.ne.tsil.ˈ os] |put$_{\text{2.sg.imp}}$ cl$_{\text{2.sg.dat}}$ cl$_{\text{3.sg.acc.m}}$ down| ‘Put yours down!’ (said to a single person) vs. *Puneți-l jos!* [ˈpu.ne.tsil.ˈ os] |put$_{\text{2.pl.imp}}$ cl$_{\text{3.sg.acc.m}}$ down| ‘Put it down!’ (said to a group of people).

A very subtle ambiguity in the current standard Romanian orthography concerns the use of hyphen: among other, the hyphen is used as sandhi marker, postverbal marker, or both (cf. also Dobrovie-Sorin and Giurgea 2013:p. 262). Appendix 1 shows some instances of hyphen usage and how it corresponds to different clitic-host structures and relations.

2.5.2. Postverbal marking

Interestingly, many languages featuring clitics employ special orthographic marking for enclitics, but not for proclitics. What is the reason for this unequal treatment? A possible explanation is the reduction of cognitive load in language processing, more precisely, the **reduction of extraneous (or extrinsic) cognitive load**, the cognitive load resulting from the way in which something is presented (cf. Sweller et al. 2019).

Due to the different positions relative to the verb and the linearity of the utterance, clitics between two verbs can be interpreted either as enclitic to the preceding verb or as proclitic to the subsequent verb. Different attempts at reducing potential ambiguities are possible: in speech, this is done with the contour of the prosodic phrase, while in writing, different specific orthographic rules are used. The postverbality of an enclitic sequence is explicitly marked in Romanian by a hyphen between the preceding verb and the clitic sequence, while in Italian and Spanish the verb and the clitic sequence are written together as a single orthographic unit.[22]

Example 1[23] and 2 illustrate my assessment of this issue. In ex. 1a, the hyphen links the noun *prietena* ‘girl-friend’ to its post-nominal possessive clitic *mi* ‘my’, an rarely used possessive construction in modern Romanian. The possessive enclitic must be evaluated with respect to the preceding noun phrase, not to the

[20] see also footnote 13

[21] used in colloquial language (cf. Dragomirescu et al. 2022:p. 245-246)

[22] This is essentially the same as using brackets in mathematics to clearly mark the scope of individual operators.

[23] adapted from Avram (1986:p. 561)

following verbal phrase.[24] In ex. 1b, the dative clitic *mi* and the accusative clitic *le* are both proclitics to the verb, and thus no marking is needed or even allowed. Ex. 2 features a similar problem in Spanish: the parsing and understanding of a series of enclitic and proclitic pronouns is eased by writing the first verb *hablar* and its enclitic *le* together as one orthographic unit – *hablarle* – so that it should not be evaluated as proclitic to the following verb, as it is the case with *lo*.

(1) a. prietena-mi le dă mere
girlfriend-my to_them give apples
'My girlfriend gives them apples.'

b. prietena mi le dă înapoi
girl-friend to_me them give back
'The girlfriend gives them back to me.'

(2) Al hablarle lo detesté.
To_the talking to_him him hated.
'When I talked to him, I hated him.'

Based on psycho-linguistic experiments on Italian single enclitics described in Finocchiaro and Caramazza (2006), Finocchiaro (2005) ponders whether the interpretation of the experiment results, namely, that enclitics pattern as affixes do with respect to gender-congruence, can be safely extended to proclitics, which are written separately from the verb, so they are "orthographically independent" (e.g., Ital. *lo pórto* |lo$_{\text{_cl.3.sg.m}}$ portare$_{\text{_1.sg.pres}}$| 'I bring it'). For Finocchiaro (2005),

> "[…] asymmetries between enclitics and proclitics are well known and appear to extend beyond superficial graphical differences. Specifically, the relation between the proclitic and the host verb appears to be less strong than the relation between the enclitic and the host verb (Benincà and Cinque 1993). Benincà and Cinque (1993) argued that the graphical difference between enclitics and proclitics corresponds to deep structural differences." (op. cit. p. 303)

However, Luraghi (2017) mentions Benincà & Cinque's (1983)[25] suggestion that

> "the fact that enclitics are often attached to their host graphically whereas proclitics are not may reflect some difference in the relation between the host and the clitic based on the direction of liaison" (op. cit. p. 189),

which is a rather reasonable suggestion.

Undeniably, there *are* differences between proclitics and enclitics due to the linearity of the utterance, the relative position to the verb as well as to the fact that both proclitics and enclitics have to be interpreted as parts of the same verbal phrase. Yet, whether differences in the orthographic rules of a language at an arbitrary point in time can be linked to deep structural differences motivated by some theory-internal assumptions is questionable.

It is useful to evaluate this claim in a more general, comparative context: in Italian, the verb and the enclitic are written together; in Romanian, they are linked by a hyphen; in Bulgarian, however, there is no difference in the orthographic representation of proclitics vs. enclitics – both are separated from the verb by a space. These observations could be interpreted to imply that in Italian, there is a stronger link between verb and enclitic than in Romanian, where there is a weaker link, and, in turn, that in Bulgarian there is the weakest or even no link between verb and enclitic. However, this is not a valid interpretation because orthographic rules are ultimately language-specific and do not reliably represent morphophonological relationships in a consistent way across languages. Indeed, every now and then they are even subject to change, independent of actual linguistic change.[26]

[24] For RWP-RWV clusters after a prosodic host, post-predicative marking is not used, as illustrated by ex. 33, where there is no hyphen between *acolo* and *mi-s*.

[25] Obviously, the mention "Benincà & Cinque (1983)" in Luraghi (2017) is a typo of the year of Benincà and Cinque (1993).

[26] see, e.g., Johnson (2005) for changes in the spelling of compounds introduced by a reform of German orthography in 1996

3. From data to model

In the previous section, I presented a detailed analysis of RWPs and RWVs as they occur in specific contexts, i.e., a static view of the data. A computational linguistic model for generating correct RWP surface forms in appropriate contexts, i.e., a procedural view, implies a set of input items (i.e., the underlying forms) and a set of rules that transform each input into the corresponding output. To construct such a model, I try to grasp patterns in language and express them in a formal way. Yet, since language is in steady change, some phenomena can be difficult to make out only by considering a synchronic perspective: the data might be somehow incomplete and/or idiosyncratic. It is essentially like looking at a painting from too close up: you can see tiny details, but not the whole picture.

By taking a step back in time as well as a step aside to some closely related language such as Aromanian, it is possible to find missing pieces to the RWP puzzle. Indeed, in the literature on Old Romanian, there is evidence for the existence of forms such as *su* and *lu* (cf. Graur 1960, Avram 1986:p. 652, or Dindelegan 2016). Consider, e.g., the instances *eu measeru-su* 'I am poor' in Dindelegan (2016:p. 169) or *nu vrea de să-lu ştie* 'he does not want anybody to know him' in Dindelegan (2016:p. 242). Moreover, a comparison between Aromanian and Romanian shows that the Aromanian form *lu* is the counterpart of the – now consonant-only – Romanian form *l*, as illustrated in Figure 2 (reproduced from Marioţeanu 1994:p. 14-15).

Figure 2: Aromanian *lu* vs. Romanian *l* in Marioţeanu (1994:p. 14-15)

As Krämer (2012) aptly notes, '[t]he existence, status and form of underlying representations have been hotly debated in phonological research', hence, it is difficult to agree on this kind of abstractions. However, the underlying representations proposed here provide a much better justification for the linguistic reality than, for instance, Dobrovie-Sorin and Giurgea's (2013:p. 266) assumption that "clitic forms are underlyingly asyllabic or syllabic", or Popescu's (2003:p. 154) assumption of unspecified underlying mora. Why is this the case? The model proposed here employs input entities evidenced in the history of Romanian and also in the closely related Aromanian, namely, items such as *su* and *lu*. Moreover, since it treats both Romanian weak pronouns and Romanian weak verbs uniformly, it offers a broader coverage of the modeled phenomena.

With this in mind, the set of *underlying items* for the model proposed here can now be established: all input items are syllabic (cf. Table 2). The constraints ruling the generation of appropriate surface forms can be derived from the description of RWP sandhi. For consonant-only RWPs/RWVs such as *l* or *s*, the model takes corresponding syllabic input[27] – *lu* or *su* – and applies the constraints derived from the data analysis. The right-most item containing the vowel *u* or *i* becomes asyllabic: *u* is deleted – cf. the optional sandhi for negation *nu* (*nu o văd* vs. *n-o văd* 'I don't see her') – while *i* becomes a palatalization gesture or a glide, depending on the context. Since the accusative always occurs after the dative, hence always in the right-most position, the syllabic input *lu* always surfaces as asyllabic, hence as a consonant. The same applies to the syllabic input *su* for the *s*-RWV forms.

This model can be implemented in computational linguistics as well. For instance, in Gerstenberger (2018), I sketch the constraints required for a computational-linguistic generation of correct RWP forms for a given context which, in turn, is couched into a general framework for linearization, the *General Linearization Model*, as proposed in Gerstenberger (2007). In this, the goal is to compare the results of this relatively simple rule-based model to the output of different statistical-based models.

[27] Already more than 60 years ago, Graur (1960:p. 847) hinted at the possibility of modeling the asyllabic *l*-RWP that way, namely using the syllabic *lu* as underlying representation.

Number	**Accusative**					**Dative**					**Verb** *a fi*	
	1p	2p	3p.m	3p.f	3p.refl	1p	2p	3p.m	3p.f	3p.refl	1p.pres	3p.pres
Sg	/mə/	/te/	/lu/	/o/	/se/	/mi/	/tsi/	/i/	/i/	/ʃi/	/su/	/i/
Pl	/ne/	/və/	/i/	/le/	/se/	/ni/	/vi/	/li/	/li/	/ʃi/	–	/su/

Table 2: Input for the surface form generation of Romanian weak pronouns and weak verb forms

4. Conclusion

In this study, I have presented an analysis of Romanian weak pronouns based on two orthogonal levels: a plain phrasal level – with the stressed verb as phrasal host and unstressed weak pronouns as phrasal clitics – and an intricate syllabic level – with the syllabic item as 'host' and the asyllabic item as 'clitic'.

Unlike the traditional descriptions hitherto, which unanimously classify pronominal *î*-prothetic forms as syllabic, I have used empirical evidence to argue for abstracting away from the prothetic *î*, and instead, classifying such forms as asyllabic. I sketched the syntactic configurations of the RWPs as well as the surface forms in which these items must, in the case of *obligatory sandhi*, or may, in the case of *optional sandhi*, occur. Through a careful examination of the orthography employed to represent RWPs, I identified different types of ambiguities that have led to an inaccurate description of syllabic postverbal RWP instances as phonological clitics by both Dobrovie-Sorin (1999a:p. 533) and Klein (2007:p. 62).

Since language is perpetually in a state of flux, it is not always possible to build a regular model for specific phenomena only from a synchronic perspective; this is made more difficult by the fuzziness of the concept 'synchronic' in terms of time frame delimitation in language description. Given this circumstance, expanding the view of the language data both historically and concerning language varieties and closely related languages such as Aromanian, I found evidence that leads to a model for RWP surface form generation without the idiosyncrasies asserted by Barbu (1999), without the assumption of unspecified underlying mora for *î*-protetic forms as in Popescu (2003:p. 154), without Klein's (2007:p. 77) employment of clusters of *î*-prothetic forms as model input, without Chereches¸' (2014:p. 56) issues with asyllabic consonantal forms lacking underlying vowels, and without Dobrovie-Sorin and Giurgea's (2013:p. 266) assumption of mixed asyllabic and syllabic underlying representations.

Finally, by providing substantiated evidence from elaborate data analyses, I have argued against a dissimilar treatment of weak pronouns occurring in preverbal as opposed to postverbal position. Hence, my answer to the question posed in the title "*How weak are Romanian clitic pronouns?*" is as follows. Since there is no crucial difference between proclitics and enclitics, there is no reason to make a distinction between weak pronouns and clitic pronouns in Romanian either.

Acknowledgements

I wish to thank Trond Trosterud and Laura Janda for their constructive discussions on this topic as well as Joshua Wilbur and Øystein Vangsnes for their thoughtful and detailed feedback on this article. Moreover, I would like to thank two anonymous reviewers for their valuable comments and suggestions on an earlier draft. Of course the usual disclaimers apply.

References

Anderson, Stephen R. 2005. *Aspects of the Theory of Clitics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199279906.001.0001.

Avram, Andrei. 1986. Sandhi Phenomena in Romanian. In *Sandhi Phenomena in the Languages of Europe*, edited by Henning Andersen, pp. 551–574. Mouton de Gruyter. https://doi.org/10.1515/9783110858532.551.

Avram, Mioara. 1997. *Gramatica pentru toţi*. Editura Humanitas, Bucureşti.

Barbu, Ana Maria. 1999. Complexul verbal. *Studii şi Cercetări Lingvistice* pp. 39–84. Available at http://dspace.bcu-iasi.ro/handle/123456789/8747.

Barbu, Roxana-Maria and Ida Toivonen. 2018. Romanian Object Clitics: Grammaticalization, agreement and lexical splits. In *Proceedings of the LFG18 Conference*, edited by Miriam Butt and Tracy Holloway King, p. 67–87. Stanford: CSLI Publications.

Benincà, Paola and Guglielmo Cinque. 1993. Su alcune differenze tra enclisi e proclisi. In *Omaggio a Gianfranco Folena (vol. 3)*, pp. 2313–2326. Editoriale Programme, Padova.

Blaszczak, Joanna, Dorota Klimek-Jankowska, and Krzysztof Migdalski (eds.) . 2015. *How Categorical are Categories?: New Approaches to the Old Questions of Noun, Verb, and Adjective*. De Gruyter Mouton. https://doi.org/10.1515/9781614514510.

Bonet, Eulàlia. 1994. The person-case constraint: A morphological approach. *MIT Working Papers in Linguistics* 0 22: 33–52.

Bonet, Eulàlia M. 1991. *Morphology after syntax: Pronominal clitics in Romance*. Ph.D. thesis, MIT. Available at http://hdl.handle.net/1721.1/13534.

Bošković, Željko. 2001. *On the Nature of the Syntax-Phonology Interface*. Emerald Group Publishing. https://doi.org/10.1163/9780585474250.

Bredemeier, Jürgen. 1976. *Strukturbeschränkungen im Rumänischen. Studien zur Syntax der prä- und postverbalen Pronomina*. TBL Verlag Gunter Narr, Tübingen.

Brown, Keith (ed.) . 2006. *Encyclopedia of Language and Linguistics*. Elsevier. Available at https://www.elsevier.com/books/encyclopedia-of-language-and-linguistics/brown/978-0-08-044299-0.

Caink, Andrew D. 2006. Clitics. In Brown (2006), pp. 491–495. https://doi.org/10.1016/B0-08-044854-2/00110-3. Available at https://www.elsevier.com/books/encyclopedia-of-language-and-linguistics/brown/978-0-08-044299-0.

Calude, Andreea S. 2001. Romanian clitics: Siding with the Serbo-Croatian or the French? *Revue roumaine de linguistique* 46 1-4: 91–104. Available at https://www.calude.net/andreea/Research.html.

Cardinaletti, Anna. 1999. Pronouns in Germanic and Romance Languages: An overview. In van Riemsdijk (1999b). https://doi.org/10.1515/9783110804010.

Cardinaletti, Anna and Michal Starke. 1999. The typology of structural deficiency: A case study of the three classes of pronouns. In van Riemsdijk (1999b). https://doi.org/10.1515/9783110804010.

Chereches, Anca. 2014. A Prosodic Analysis of Romanian Pronominal Clitics. In *University of Pennsylvania Working Papers in Linguistics*, vol. 20. Available at https://repository.upenn.edu/pwpl/vol20/iss1/7/.

Di Sciullo, Anna Maria and Edwin Williams. 1987. *On the Definition of Word*. MIT Press, Cambridge (Mass.). Available at https://mitpress.mit.edu/books/definition-word.

Dindelegan, Gabriela Pană (ed.) . 2013. *The Grammar of Romanian*. Oxford University Press. Available at https://global.oup.com/academic/product/the-grammar-of-romanian-9780199644926.

Dindelegan, Gabriela Pană (ed.) . 2016. *The Syntax of Old Romanian*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198712350.001.0001.

Dindelegan, Gabriela Pană, Adnana Boioc Apintei, and Blanca Croitor (eds.) . 2019. *Variație diacronică și diatopică: Note gramaticale*. Editura Universităţii din Bucureşti. Available at https://editura-unibuc.ro/en/magazin/filologie/limba-romana/variatie-diacronica-si-diatopica-note-gramaticale/.

Dobrovie-Sorin, Carmen. 1994. *The Syntax of Romanian: Comparative Studies in Romance*. Mouton de Gruyter. https://doi.org/10.1515/9783110886597.

Dobrovie-Sorin, Carmen. 1999a. Clitics across categories: The case of Romanian. In van Riemsdijk (1999b).

https://doi.org/10.1515/9783110804010.
Dobrovie-Sorin, Carmen. 1999b. The typology of pronouns and the distinction between syntax and morphophonology. In van Riemsdijk (1999b). https://doi.org/10.1515/9783110804010.
Dobrovie-Sorin, Carmen and Ion Giurgea (eds.) . 2013. *A Reference Grammar of Romanian*, vol. 1: The noun phrase. John Benjamins. https://doi.org/10.1075/la.207.
Dragomirescu, Adina, Alexandru Nicolae, and Rodica Zafiu. 2022. The loss of analyticity in the history of Romanian verbal morphology. In Ledgeway et al. (2022). https://doi.org/10.1093/oso/9780198870807.003.0010.
Finocchiaro, Chiara. 2005. Psychological evidence on the status of Romance clitics. *Italian Journal of Linguistics* 17 2: 291–310. Available at https://www.italian-journal-linguistics.com/2005-2/.
Finocchiaro, Chiara and Alfonso Caramazza. 2006. The production of pronominal clitics: Implications for theories of lexical access. *Language and Cognitive Processes* 21 1-3: 141–180. https://doi.org/10.1080/01690960400001887.
Firică, Camelia. 2009. The phonetic or the etymological principle in Romanian orthography? *Govor* 26 1: 53–62. Available at https://hrcak.srce.hr/165892.
Franco, Ludovico and Paolo Lorusso (eds.) . 2019. *Linguistic Variation: Structure and Interpretation*. De Gruyter Mouton. https://doi.org/10.1515/9781501505201.
Franks, S. and T.H. King. 2000. *A Handbook of Slavic Clitics*. Oxford Studies in Comparative Syntax. Oxford University Press. Available at https://global.oup.com/academic/product/a-handbook-of-slavic-clitics-9780195135886.
Gerlach, B. and J. Grijzenhout (eds.) . 2001. *Clitics in Phonology, Morphology and Syntax*. John Benjamins. https://doi.org/10.1075/la.36.
Gerlach, Birgit. 2002. *Clitics Between Syntax and Lexicon*. John Benjamins. https://doi.org/10.1075/la.51.
Gerstenberger, Ciprian-Virgil. 2007. A mereology-based general linearization model for surface realization. In *Proceedings of EUROLAN 2007*. University of Iaşi, Romania. Available at https://www.researchgate.net/publication/233951937_A_mereology-based_general_linearization_model_for_surface_realization.
Gerstenberger, Ciprian-Virgil. 2018. A grammar for romanian weak pronoun generation. In *Languages at the Crossroads: Training, Accreditation and Context of Use. Proceedings of the 35th Edition of the International Conference of the Spanish Association of Applied Linguistics*, edited by Francisco Javier Díaz Pérez and M.ª Águeda Moreno Moreno, pp. 215–226. University of Jaén. Available at https://www.researchgate.net/publication/332671041_A_Grammar_for_Romanian_weak_pronoun_generation.
Giurgea, Ion. 2013. L'origine des clitiques roumains de 3e personne pluriel datif et de 1ere personne pluriel datif-accusatif. *Revue roumaine de linguistique* LVIII 2: 131–143.
Graur, Al. 1960. Observaţii asupra sinerezei în romîneşte [Observations concerning synaeresis in Romanian]. *Studii şi Cercetări Lingvistice* .
Guţu-Romalo, Valeria (ed.) . 2008. *Gramatica limbii române*. Editura Academiei Române.
Harris, Alice C. 2002. *Endoclitics and the Origins of Udi Morphosyntax*. Oxford University Press. Available at https://global.oup.com/academic/product/endoclitics-and-the-origins-of-udi-morphosyntax-9780199246335.
Haspelmath, Martin. 2015. Defining vs. diagnosing linguistic categories: A case study of clitic phenomena. In Blaszczak et al. (2015), pp. 273–304. https://doi.org/10.1515/9781614514510-009.
Haspelmath, Martin. 2022. Types of clitics in the world's languages. In preparation; available at https://www.academia.edu/s/52c1b938dc.
Hill, Virginia and Gabriela Alboiu. 2016. *Verb Movement and Clause Structure in Old Romanian*. Oxford Studies in Diachronic and Historical Linguistics. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780198736509.001.0001.
Iliescu, Maria. 1975. Pentru o sistematizare a predării pronumelui personal neaccentuat românesc (la studenţii străini). *Limba Română* 24: 51–62. Available at https://www.diacronia.ro/en/indexing/details/A11324.

Jianu, Maria-Magdalena. 2013. Observations upon the Clitics of the Dative Case in Romanian. *International Journal of Communication Research* 3 2: 117.

Johnson, Sally. 2005. *Spelling Trouble?: Language, Ideology and the Reform of German Orthography*. Multilingual Matters. Available at https://www.multilingual-matters.com/page/detail/Spelling-Trouble-Language-Ideology-and-the-Reform-of-German-Orthography/?k=9781853597848.

Kabatek, Johannes, Philipp Obrist, and Albert Wall (eds.) . 2021. *Differential Object Marking in Romance: The third wave*. De Gruyter. https://doi.org/10.1515/9783110716207.

Klavans, Judith L. (ed.) . 1995. *On Clitics and Cliticization: The Interaction of Morphology, Phonology, and Syntax*. Routledge. https://doi.org/10.4324/9780429442636.

Klein, Udo-Michael. 2007. *Encoding of argument structure in Romanian and SiSwati*. Ph.D. thesis, University of London. Available at http://hdl.handle.net/11858/00-001M-0000-0012-8EEC-D.

Krämer, Martin. 2012. *Underlying Representations*. Cambridge University Press. https://doi.org/10.1017/CBO9780511978821.

Ledgeway, Adam, John Charles Smith, and Nigel Vincent (eds.) . 2022. *Periphrasis and Inflexion in Diachrony: A View from Romance*. Oxford Studies in Diachronic and Historical Linguistics Series. Oxford University Press. https://doi.org/10.1093/oso/9780198870807.001.0001.

Legendre, Géraldine. 2001. Positioning Romanian verbal clitics at PF. In Gerlach and Grijzenhout (2001). https://doi.org/10.1075/la.36.10leg.

Lombard, Alf. 1976. Le î prosthétique du roumain. *Acta Societatis linguisticae Upsaliensis* 2 5.

Luraghi, Silvia. 2017. Clitics. In Luraghi and Parodi (2017), chap. 11, p. 165–193. https://doi.org/10.5040/9781472542090.

Luraghi, Silvia and Claudia Parodi (eds.) . 2017. *The Bloomsbury Companion to Syntax*. Continuum. https://doi.org/10.5040/9781472542090.

Maiden, Martin, Adina Dragomirescu, Gabriela Pană Dindelegan, Oana Uţă, and Rodica Zafiu. 2021. *The Oxford History of Romanian Morphology*. Oxford University Press. https://doi.org/10.1093/oso/9780198829485.001.0001.

Manzini, M. Rita. 2014. Grammatical categories: Strong and weak pronouns in Romance. *Lingua* 150: 171–201. https://doi.org/10.1016/j.lingua.2014.07.001.

Marioţeanu, Matilda Caragiu (ed.) . 1994. *Di nuntru şi-di nafoarâ: Stihuri armâneşti*. Cartea Românească.

Mišeska Tomić, Olga. 2006. *Balkan Sprachbund Morpho-Syntactic Features*. Springer. https://doi.org/10.1007/1-4020-4488-7.

Monachesi, Paola. 2001. Clitic placement in the Romanian verbal complex. In Gerlach and Grijzenhout (2001). https://doi.org/10.1075/la.36.11mon.

Monachesi, Paola. 2005. *The Verbal Complex in Romance: A Case Study in Grammatical Interfaces*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199274758.001.0001.

Montreuil, Jean-Pierre Y. (ed.) . 2006. *New Perspectives on Romance Linguistics*. John Benjamins. https://doi.org/10.1075/cilt.276.

Nastasenco, O. A. 1997. *Gramatica limbii române în tabele*. Editura VIRT, Chişinău.

Nevins, Andrew and Oana Săvescu. 2008. An Apparent 'Number Case Constraint' in Romanian: The Role of Syncretism. *Romance Linguistics 2008. Selected Papers from the 36th Linguistic Symposium on Romance Languages* https://doi.org/10.1075/cilt.313.17nev.

Ordóñez, Francisco and Lori Repetti. 2006. Stressed Enclitics? In Montreuil (2006). https://doi.org/10.1075/cilt.276.13ord.

Pescarini, Diego. 2018. Stressed enclitics are not weak pronouns: A plea for allomorphy. In *Romance Languages and Linguistic Theory 14: Selected papers from the 46th Linguistic Symposium on Romance Languages (LSRL)*. Stony Brook. https://doi.org/10.1075/rllt.14.13pes.

Pescarini, Diego. 2019. An emergentist view on functional classes. In Franco and Lorusso (2019), pp. 531–560. https://doi.org/10.1515/9781501505201-027.

Popescu, Alexandra. 2000. The morphophonology of the Romanian clitic sequence. In *Lingua*, vol. 110, pp. 773–799. Elsevier. https://doi.org/10.1016/S0024-3841(00)00016-4.

Popescu, Alexandra. 2003. *Morphophonologische Phänomene des Rumänischen*. Ph.D. thesis, University of Düsseldorf. Available at https://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=3187.

van Riemsdijk, Henk C. 1999a. Clitics: A state-of-the-art report. In van Riemsdijk (1999b). https://doi.org/10.1515/9783110804010.1.

van Riemsdijk, Henk C. (ed.) . 1999b. *Clitics in the Languages of Europe*. Mouton de Gruyter. https://doi.org/10.1515/9783110804010.

Rosetti, Alexandru. 1986. *Istoria limbii române*. Editura ştiinţifică şi enciclopedică.

Rouveret, Alain. 1999. Clitics, subjects and tense in European Portuguese. In van Riemsdijk (1999b). https://doi.org/10.1515/9783110804010.

Sasaki, Kan and Daniela Căluianu. 2000. An Optimality Theoretic Account for the Distribution of Pronominal Clitics in Romanian. Tech. rep., University of Tsukuba. Available at https://www.academia.edu/33668351/An_optimality_theoretic_account_for_the_distribution_of_pronominal_clitics_in_Romanian.

Somesfalean, Stanca. 2007. *On the Form and Interpretation of Clitics*. Ph.D. thesis, Université du Québec à Montréal. Available at https://archipel.uqam.ca/9628/.

Spencer, Andrew and Ana Luís. 2012. *Clitics: An Introduction*. Cambridge University Press.

Săvescu Ciucivara, Oana. 2009. *A Syntactic Analysis of Pronominal Clitic Clusters in Romance - The view from Romanian*. Ph.D. thesis, New York University. Available at https://www.proquest.com/docview/304954033.

Sweller, John, Jeroen J. G. van Merriënboer, and Fred Paas. 2019. Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review* 31 2: 261–292. https://doi.org/10.1007/s10648-019-09465-5.

Tigău, Alina. 2021. Differential Object Marking in Romanian and Spanish: A contrastive analysis between differentially marked and unmarked direct objects. In Kabatek et al. (2021), pp. 173–212. https://doi.org/10.1515/9783110716207-007.

Zafiu, Rodica. 2019. Schimbarea morfologică în paradigma de prezent a verbului „a fi": formele bănăţene „mi(-)s", „ni(-)s", „vi(-)s". In Dindelegan et al. (2019). Available at https://editura-unibuc.ro/en/magazin/filologie/limba-romana/variatie-diacronica-si-diatopica-note-gramaticale/.

Zwicky, Arnold. 1977. On clitics. *Bloomington, IN: Indiana University Linguistics Club* Available at https://arnoldzwicky.org/about/papers/.

Zwicky, Arnold. 1985. Clitics and Particles. *Language* 61 2: 283–305. https://doi.org/10.2307/414146.

Zwicky, Arnold M. and Geoffrey K. Pullum. 1983. Cliticization vs. inflection: English n't. *Language* 59 3: 502–513. https://doi.org/10.2307/413900.

Appendix 1: Hyphen and syllabicity disambiguation for pre- and post-verbal Romanian weak pronouns

Example	**[aɪ pʰiː eɪ]**	***Orthography***
Gloss	**Disambiguation**	**'Translation'**
(3) *î -l cumperi* î_host cl_3.sg.acc.m buy_2.sg.pres	[ɨl.ˈcum.perⁱ] [î] (l) cumperi	*Îl cumperi.* 'You buy it.'
(4) *cumpără -l* buy_2.sg.imp cl_3.sg.acc.m	[ˈcum.pə.rəl] [cumpără] — (l)	*Cumpără-l!* 'Buy it!'
(5) *ți -l cumperi* cl_2.sg.dat cl_3.sg.acc.m buy_2.sg.pres	[tsil.ˈcum.perⁱ] [ți] (l) cumperi	*Ți-l cumperi.* 'You buy it.'
(6) *cumpără -ți -l* buy_2.sg.imp cl_2.sg.dat cl_3.sg.acc.m	[ˈcum.pə.rə.tsil] cumpără — [ți] — (l)	*Cumpără-ți-l!* 'Buy it!'
(7) *ți- o cumperi* cl_2.sg.dat cl_3.sg.acc.f buy_2.sg.pres	[tsi̯o.ˈcum.perⁱ] (ți) [o] cumperi	*Ți-o cumperi.* 'You buy it.'
(8) *cumpără -ți -o* buy_2.sg.imp cl_2.sg.dat cl_3.sg.acc.f	[ˈcum.pə.rə.tsi̯o] cumpără — (ți) — [o]	*Cumpără-ți-o!* 'Buy it!'
(9) *să -ți cumperi cartea* that cl_2.sg.dat buy_2.sg.pres book_def	[sətsⁱ.ˈcum.perⁱ.ˈcar.te̯a] [să] (ți) cumperi cartea	*Să-ți cumperi cartea!* 'Buy the book!'
(10) *cumpără -ți cartea* buy_2.sg.imp cl_2.sg.dat book_def	[ˈcum.pə.rətsⁱ.ˈcar.te̯a] [cumpără] — (ți) cartea	*Cumpără-ți cartea!* 'Buy the book!'
(11) *ți- ai cumpărat cartea* cl_2.sg.dat have_2.sg.pres buy_past.part book_def	[tsi̯ai̯.cum.pə.ˈrat.ˈcar.te̯a] (ți) [ai] cumpărat cartea	*Ți-ai cumpărat cartea.* 'You've bought the book.'
(12) *te duci acasă* cl_2.sg.acc.refl carry_2.sg.pres home	[te.ˈdutʃⁱ.a.ˈca.sə] [te] duci acasă	*Te duci acasă.* 'You go home.'
(13) *du -te acasă* carry_2.sg.imp cl_2.sg.acc.refl home	[ˈdu.te.a.ˈca.sə] du — [te] acasă	*Du-te acasă!* 'Go home!'
(14) *du -te -acasă* carry_2.sg.imp cl_2.sg.acc.refl home	[ˈdu.te̯a.ˈca.sə] du — (te) [acasă]	*Du-te-acasă!* 'Go home!'
(15) *mi le faci acum* cl_1.sg.dat cl_2.pl.acc.f do_2.sg.pres now	[mi.le.ˈfatʃⁱ.a.ˈcum] [mi] [le] faci acum	*Mi le faci acum.* 'You do them for me now.'
(16) *fă -mi -le acum* do_2.sg.imp cl_1.sg.dat cl_2.pl.acc.f now	[ˈfə.mi.le.a.ˈcum] fă — [mi] — [le] acum	*Fă-mi-le acum!* 'Do them for me now!'
(17) *fă -mi -le -acum* do_2.sg.imp cl_1.sg.dat cl_2.pl.acc.f now	[ˈfə.mi.le̯a.ˈcum] fă — [mi] — (le) [acum]	*Fă-mi-le-acum!* 'Do them for me now!'

Legend: [xx] → **syllabic weak pronoun** (xx) → **asyllabic weak pronoun**
[xx] → **syllabic host** [[xx]] → **syllabic weak pronoun and syllabic host**
[xx] (yy) → **sandhi marker** xx — [yy] → **postverbal marker** [xx] — (yy) → **sandhi and postverbal marker**

Appendix 2: Examples of analyzed Romanian weak pronouns

Glossed examples	**Orthography**	**IPA**	**Translation**
(18) dă -l give$_{_2.sg.imp}$ cl$_{_3.sg.acc.m}$	*Dă-l!*	[ˈdəl]	‘Give it!’
(19) a -l da to cl$_{_3.sg.acc.m}$ give$_{_inf}$	*a-l da*	[al.ˈda]	‘to give it’
(20) a î -l da to î$_{_host}$ cl$_{_3.sg.acc.m}$ give$_{_inf}$	*a îl da*	[a.ɨl.ˈda]	‘to give it’
(21) dându -l give$_{_ger}$ cl$_{_3.sg.acc.m}$	*dându-l*	[ˈdɨn.dul]	‘giving it’
(22) nu -l da not cl$_{_3.sg.acc.m}$ give$_{_inf}$	*Nu-l da!*	[ˈnul.ˈda]	‘Don’t give it!’
(23) nu î -l da! not î$_{_host}$ cl$_{_3.sg.acc.m}$ give$_{_inf}$	*Nu îl da!*	[ˈnu.ɨl.ˈda]	‘Don’t give it!’
(24) mi -l dai cl$_{_1.sg.dat}$ cl$_{_3.sg.acc.m}$ give$_{_2.sg.pres}$	*Mi-l dai.*	[mil.ˈdaɪ̯]	‘You give it to me.’
(25) dă -le give$_{_2.sg.imp}$ cl$_{_3.pl.acc.f}$	*Dă-le!*	[ˈdə.le]	‘Give them!’
(26) mi le dai acum cl$_{_1.sg.dat}$ cl$_{_3.pl.acc.f}$ give$_{_2.sg.pres}$ now	*Mi le dai acum.*	[mi.le.ˈdaɪ̯.a.ˈcum]	‘You give them to me now.’
(27) dă -mi -le acum give$_{_2.sg.imp}$ cl$_{_1.sg.dat}$ cl$_{_3.pl.acc.f}$ now	*Dă-mi-le acum!*	[ˈdə.mi.le.a.ˈcum]	‘Give them to me now!’
(28) dă -mi -le -acum give$_{_2.sg.imp}$ cl$_{_1.sg.dat}$ cl$_{_3.pl.acc.f}$ now	*Dă-mi-le-acum!*	[ˈdə.mi.le̯a.ˈcum]	‘Give them to me now!’
(29) se -ntâmplă cl$_{_3.acc.refl}$ happen$_{_3.sg.pres}$	*Se-ntâmplă.*	[sen.ˈtɨm.plə]	‘It happens.’
(30) se întâmplă cl$_{_3.acc.refl}$ happen$_{_3.sg.pres}$	*Se întâmplă.*	[se.ɨn.ˈtɨm.plə]	‘It happens.’
(31) î -l pot vedea î$_{_host}$ cl$_{_3.sg.acc.m}$ can$_{_1.sg.pres}$ see$_{_inf}$	*Îl pot vedea.*	[ɨl.ˈpot.veˈde̯a]	‘I can see him/it.’
(32) copiii mi -s acolo child$_{_pl.def}$ cl$_{_1.sg.dat}$ be$_{_3.pl.pres}$ there	*Copiii mi-s acolo.*	[co.ˈpi.i̯i.mis.a.ˈco.lo]	‘My children are there.’
(33) acolo mi -s copiii there cl$_{_1.sg.dat}$ be$_{_3.pl.pres}$ child$_{_pl.def}$	*Acolo mi-s copiii!*	[a.ˈco.lo.mis.co.ˈpi.i̯i]	‘There are my children!’
(34) mi -o dai cl$_{_1.sg.dat}$ cl$_{_3.sg.acc.f}$ give$_{_2.sg.pres}$	*Mi-o dai.*	[mi̯o.ˈdaɪ̯]	‘You give her/it to me.’
(35) mi -ai dat -o cl$_{_1.sg.dat}$ have$_{_2.sg.pres}$ given cl$_{_3.sg.acc.f}$	*Mi-ai dat-o.*	[mi̯aɪ̯.ˈda.to]	‘You have given her/it to me.’
(36) te -aș vedea sănătos cl$_{_2.sg.acc}$ have$_{_1.sg.cond}$ see$_{_inf}$ healthy	*Te-aș vedea sănătos!*	[te̯aʃ.ve.ˈde̯a.sə.nə.ˈtos]	‘May I see you healthy!’
(37) vedea -te -aș sănătos see$_{_inf}$ cl$_{_2.sg.acc}$ have$_{_1.sg.cond}$ healthy	*Vedea-te-aș sănătos!*	[ve.ˈde̯a.te̯aʃ.sə.nə.ˈtos]	‘May I see you healthy!’
(38) vedea -o -aș moartă see$_{_inf}$ cl$_{_3.sg.acc.f}$ have$_{_1.sg.cond}$ dead	*Vedea-o-aș moartă!*	[ve.ˈde̯a.o̯aʃ.ˈmo̯ar.tə]	‘May I see her dead!’
(39) mi l- ai dat cl$_{_1.sg.dat}$ cl$_{_3.sg.acc.m}$ have$_{_2.sg.pres}$ given	*Mi l-ai dat.*	[mi.laɪ̯.ˈdat]	‘You have given it to me.’
(40) că -mi dai mere that cl$_{_1.sg.dat}$ give$_{_2.sg.pres}$ apples	*că-mi dai mere*	[cəmʲ.ˈdaɪ̯.ˈme.re]	‘that you give me apples’
(41) că î -mi dai mere that î$_{_host}$ cl$_{_1.sg.dat}$ give$_{_2.sg.pres}$ apples	*că îmi dai mere*	[cə.ɨmʲ.ˈdaɪ̯.ˈme.re]	‘that you give me apples’
(42) mi- aduci mere cl$_{_1.sg.dat}$ bring$_{_2.sg.pres}$ apples	*Mi-aduci mere.*	[mi̯a.ˈdutʃʲ.ˈme.re]	‘You bring me apples.’
(43) î -mi aduci mere. î$_{_host}$ cl$_{_1.sg.dat}$ bring$_{_2.sg.pres}$ apples	*Îmi aduci mere.*	[ɨmʲ.a.ˈdutʃʲ.ˈme.re]	‘You bring me apples.’
(44) roua mi -mbată inima dew$_{_def}$ cl$_{_1.sg.dat}$ makes drunk heart$_{_def}$	*Roua mi-mbată inima.*	[ˈro.wa.mim.ˈba.tə.ˈi.ni.ma]	‘The dew makes my heart drunk.’
(45) roua î -mi îmbată inima. dew$_{_def}$ î$_{_host}$ cl$_{_1.sg.dat}$ makes drunk heart$_{_def}$	*Roua îmi îmbată inima.*	[ˈro.wa.ɨmʲ.ɨm.ˈba.tə.ˈi.ni.ma]	‘The dew makes my heart drunk.’

Mari morpheme order revisited: a corpus-based analysis

Luan Hammer and Jeremy Bradley
University of Vienna

Abstract

Morpheme order in Mari declension has been extensively studied in the past, but attempts to explain the large amounts of alternation found here have been constrained by the difficulty of accessing sufficient data to properly elucidate the complexities in this domain. The paper at hand examines the prospect of using the Corpus of Literary Mari, created by an international workgroup around Trond Trosterud and his colleagues and hosted by Giellatekno, and other recently published resources on Mari to efficiently access vast amounts of data to quantitatively study this subject in a manner that had not previously been possible.

Keywords: Corpus, Mari language, suffix order, possessive suffixes

1. Mission statement

The unusually large freedom the Mari language(s) of European Russia afford their speakers as regards the arrangement of case (CX), possessive (PX), and number suffixes (NX) in nominal morphology have long puzzled and perplexed scholars and language learners alike. While not all six possible arrangements of these are permissible, substantial variation can be encountered:

(1) Meadow Mari (Corpus of Literary Mari)

a. *joltaš-em-βlak-lan*	b. *pire-βlak-et-lan*	c. *joča-βlak-lan-že*
friend-1SG-PL-DAT	wolf-PL-2SG-DAT	child-PL-DAT-3SG
‘to my friends’	‘to your wolves’	‘to his/her/theirSG children’
(PX-NX-CX)	(NX-PX-CX)	(NX-CX-PX)

Reference materials (including those co-authored by authors of this paper) will often try to handwave this alternation away by labelling competing forms as equivalent, but linguists and language teachers describing multiple forms as equivalent are generally tacitly admitting (intentionally or not) that they simply do not understand the factors governing the alternation – be they on the level of morphosyntax, semantics, information structure, or dialectology.

Jorma Luutonen’s 1997 dissertation “The variation of morpheme order in Mari declension” is an impressive attempt to elucidate this question. It gives an exhaustive and satisfying overview on the constraints in this domain and gives a comprehensive quantitative overview of the respective frequencies of different arrangements, though the factors governing this variation remained elusive – as they do today.

More than two decades later, this question deserves re-examination due to revolutions in the quantitative study of morphologically rich minority languages, driven forward especially by the Tromsø-based Giellatekno work group under the leadership of Trond Trosterud. While Giellatekno’s foundational mission pertains to guaranteeing language technology for the Saami languages, its open infrastructure – and especially the broad horizon of its founder and leader – has allowed other language and scholarly communities to profit from its framework. For Mari, this has meant the creation of the Corpus of Literary Mari (henceforth CLM) by a work group including Trond Trosterud, Jorma Luutonen, and many others, a first release of which, covering 57.38 million tokens of Meadow Mari texts, was soft launched in December 2020 at gtweb.uit.no/u_korp/?mode=mhr, with meta information on the project and instructions published at corpus.mari-language.com. Texts are taken from non-fiction texts, fiction texts, legal texts, scientific texts, news texts and Wikipedia texts, and they span over a century of Mari literacy (1912–2018). In this paper, we will explore the prospect of using new corpus infrastructures as novel tools to revisit enigmas

 Nordlyd 46.1: 59–74, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd
https://doi.org/10.7557/12.6373

such as the alternation in suffix order in Mari declension. Given that this should only be understood as a pilot study, we are currently restricting ourselves to the dominant Meadow Mari literary norm, which is better covered by the corpus at this point than the secondary literary norm, Hill Mari.

2. Variation in suffix order in the Meadow Mari nominal paradigm

Mari is a Uralic language, or cluster of languages, spoken by several hundred thousand speakers mainly in the titular Republic of Mari El and by a sizable diaspora in the Ural Mountains. Most reference materials and dialectal surveys distinguish between four dialect groups (Meadow, Hill, Eastern, Northwestern); two literary norms (Meadow, Hill) are widely used today. There is an ongoing debate among native speakers and linguists on the status of Hill Mari, if it should be considered a language or a dialect – cf. Trosterud & Alikov (1994) for a discussion of this problem and a comparison with the discourse surrounding Bokmål and Nynorsk in Norway. This chapter will restrict itself to the Meadow Mari literary norm; all language data in this chapter is in Meadow Mari.

Mari is, together with neighbouring Uralic and Turkic languages, generally classified as a member of the *Volga-Kama Sprachbund* in which a large amount of linguistic convergence can be observed. Mari shows not just lexical but also significant structural influence from two Turkic languages, Chuvash and Tatar. The dominant contact language today is, unsurprisingly, Russian; Russian structural influence on Mari is salient as well.

As is typical of the languages of this region, Mari morphology is rich and highly concatenative, i.e., morpheme boundaries can almost always be easily and unambiguously identified; portmanteau morphemes are virtually non-existent. Mari morphosyntax does not have dual forms. Nominal stems can take number suffixes (e.g., *olma* ‘apple’ > *olma-βlak* apple-PL ‘apples’, see below about other plural markers), case suffixes (e.g., *olma-lan* apple-DAT ‘to the apple’), possessive suffixes (e.g., *olma-na* apple-1PL ‘our apple’), and word-final clitics (e.g., *olma=t* apple=ADD ‘also the apple’). It should be noted that the 3SG and 2SG possessive suffixes have secondary usages where they serve more as determining elements indicating a topic, a contrast, or that something is part of a salient whole (cf. Simonenko 2014, Riese et al. 2019: 60); in this function the possessive suffixes behave like clitics and generally appear in word-final position; they can also be used in combination with other possessive suffixes in possessive function (e.g. *ačа-t-še* father-2SG-3SG ‘as for your father’, CML).

The permissible combinations of suffixes belonging to these types (excluding clitics) and alternation encountered in the language are illustrated in Tables 1 and 2, with Table 1 showing the Mari noun *olma* ‘apple’ in all cases (singular and plural) and Table 2 adding the possessive suffix of the second person singular to it. This data was generated using paradigm.mari-language.com, with some forms added that, thanks to the Corpus of Literary Mari, we can assume to exist. It should be noted that the permissible combinations and their frequencies differ between different possessive suffixes, i.e., one should not extrapolate from the possessive suffix 2SG to other possessive suffixes.

Case	Singular	Plural
Nominative	olma	olma-βlak
Genitive	olma-n	olma-βlak-ə̂n
Dative	olma-lan	olma-βlak-lan
Accusative	olma-m	olma-βlak-ə̂m
Comparative[1]	olma-la	olma-βlak-la
Comitative	olma-ge	olma-βlak-ge
Inessive	olma-šte	olma-βlak-ə̂šte
Illative	olma-š(ke)	olma-βlak-ə̂š(ke)[2]
Lative	olma-š	olma-βlak-eš

Table 1: Paradigm of the Mari noun *olma* ‘apple’

Case	Singular	Plural
Nominative	olma-t	olma-t-βlak ~ olma-βlak-et
Genitive	olma-t-ə̂n	olma-t-βlak-ə̂n ~ olma-βlak-et-ə̂n
Dative	olma-t-lan ~ olma-lan-et	olma-t-βlak-lan ~ olma-βlak-et-lan ~ olma-βlak-lan-et
Accusative	olma-t-ə̂m	olma-t-βlak-ə̂m ~ olma-βlak-et-ə̂m
Comparative	olma-t-la ~ olma-la-t	olma-t-βlak-la ~ olma-βlak-et-la ~ olma-βlak-la-t
Comitative	olma-t-ge	olma-t-βlak-ge ~ olma-βlak-et-ge
Inessive	olma-št-et	olma-βlak-ə̂št-et ~ olma-t-βlak-ə̂šte
Illative	olma-šk-et	olma-βlak-ə̂šk-et ~ olma-t-βlak-ə̂š(ke)
Lative	olma-š-et	olma-βlak-eš-et ~ olma-t-βlak-eš

Table 2: Paradigm of the Mari noun *olma* ‘apple’ with possessive suffix 2Sg

Different suffix orders depending on the specific case, as encountered here, is not exclusive to Mari; it can also be encountered for example in Moksha Mordvin, Permic, and Southern Samoyedic (Honti 1995, Tauli

[1] The comparative case denotes a likeness or similarity (e.g., *olma-la* apple-CMPR ‘like an apple’) and is not to be confused with the comparative degree (e.g. *joškargə̂-rak* red-COMP degree ‘redder’). Presumably motivated by this terminological confusion, other sources refer to this case as ‘equitative’, ‘similative’, or ‘modal’.

[2] The illative suffix has a short form and a long form; the short form can only be used if the illative suffix is the final suffix attached to a stem.

1953). What is unusual, however, are the many degrees of freedom, esp. in the dative and comparative cases.

While Mari generally uses plural markers relatively sparingly, with formally singular forms being used when plural semantics can be assumed from context (cf. Riese et al. 2019: 55–57), four different plural markers can be used in literary Meadow Mari. In addition to the plural suffix *-βlak* that is shown in the tables, Mari has plural forms in *-šamə̑č*, *-mə̑t*, and *-la*. The distribution of *-βlak* and *-šamə̑č* seems to be governed by dialectal distribution, with both forms used in literary language today. The other plural markers have clearly defined distinct functions: *-mə̑t* is an associative/heterogenous plural marker 'x and those associated with x, x & co.', e.g., *ača-mə̑t* father-PL.ASS 'father and those with him'. This suffix can be preceded by possessive suffixes (e.g., *ača-t-mə̑t* father-2SG-PL.ASS 'your father and those with him') but only possessive suffixes used in non-possessive functions can follow it (*Irina-mə̑t-še* Irina-PL.ASS-3SG 'as for Irina and co.', *ruš spekul'ant-mə̑t-et* Russian profiteer-PL.ASS-2SG 'that Russian profiteer and his buddies', CLM). The plural in *-la* on the other hand is primarily usually used with inanimate nouns in a local meaning in combination with local case endings and spatial adverbs (e.g., *pört-la-šte* house-PL.LOC-INE 'in the houses'). While it is uncommon for this plural suffix to co-occur with possessive suffixes outside of their non-possessive functions (Riese et al. 2019: 59), examples can be found in which the possessive suffix follows the plural and case suffixes (*ojlə̑maš-la-št-em* story-PL.LOC-INE-1SG 'in my stories', CLM).

Some fundamental observations can be made regarding restrictions in the ordering of suffixes, irrespective the large variation that can still be observed, as illustrated in the tables:

1. Case suffixes always follow number suffixes.
2. Possessive suffixes can either precede or follow number suffixes.
3. The genitive, accusative, and comitative suffixes occur in the final position; only clitics and possessive suffixes used as clitics in a non-possessive function can follow them (e.g., *mə̑j-ən-že*[3] 1SG-GEN-3SG 'as for mine', *βolgə̑də̑-m-žo* light-ACC-3SG 'as for the light', CML)
4. The local case markers – inessive, illative, lative – precede possessive suffixes. A rare exception is when a possessive suffix precedes the number suffix; in this case, the local case suffix can be encountered in final position technically following the possessive suffix (e.g., *užaš-ə̑že-βlak-ə̑šte* part-3SG-PL-INE 'in its parts', CML).
5. In the dative and comparative case, possessive suffixes can either precede or follow case suffixes.

As regards the variation, Jorma Luutonen's 1997 dissertation provides statistical data on the frequency of different variants. Elina Guseva and Philipp Weisser touched on the topic in an article where they explain the different morpheme order of structural and local cases in Meadow Mari by postsyntactic operations (Guseva & Weisser 2018). In general, existing case studies on the order of morphemes for specific languages mainly concern verbs (e.g., Rice 2000 for Athapaskan languages, Caballero 2010 for Choguita Raràmuri) and apart from the aforementioned works by Luutonen and Guseva & Weisser we are not aware of any case studies on suffix order variation that are dedicated specifically to nouns.

In this paper, we will revisit the variation described in the literature using new corpora of Mari, on the one hand to re-examine Luutonen's findings, on the other hand to look for possible additional information and explanations. The main research questions are: Which of the variants are used the most frequently? Are there differences between the frequencies of the use of the variants between different possessive suffixes or cases? Did the distribution of the variants change over time? What motivates the usage of one or another suffix order variant?

[3] It should be noted that the placement of the possessive suffix third person singular, here realized as *-že*, mirrors the placement of the homophonous Russian clitic *že*, which yields the possibility of a structural influence from Russian. However, other realizations of the possessive suffix, such as *-žo* in the following example – it can thus be reasonably assumed that Mari speakers perceive these as possessive suffixes of the third person singular, even if the range of usages has been influenced by Russian.

3. A closer look at Mari corpora

The Corpus of Literary Mari (CLM) was already introduced in the introduction. The second resource we have utilized are the Meadow Mari corpora compiled by Timofey Arkhangelskiy and his colleagues as part of a larger endeavour to create corpora for languages of the Volga-Kama Region (henceforth VK Corpora; other corpora as of the moment of this writing published by this work group cover Erzya, Moksha, Udmurt, Komi-Zyrian; a general overview can be found at volgakama.web-corpora.net); this resource consists of two basic corpora: a general corpus of literary texts (henceforth VK-MAIN), and the Social Media Corpus (henceforth VK-SMC) consisting mainly of texts curated from the platform *vkontakte* and also including Russian texts and ample amounts of code mixing (for further information on the development of the Social Media Corpora see Arkhangelskiy 2019). Table 3 shows the token counts, as of the compilation of this paper, of the three main resources at our disposal.

CLM	gtweb.uit.no/u_korp/?mode=mhr	57.38 million
VK-Main	meadow-mari.web-corpora.net/meadow-mari_corpus	5.53 million
VK-SMC (Mari only)	meadow-mari.web-corpora.net/meadow-mari_social_media	3.59 million

Table 3: Tokens in the corpora used as of 20 October 2021

Table 4 summarizes some main structural differences between CLM on the one hand, and the VK Corpora on the other.

CLM	VK Corpora
Mari lemma forms are provided as part of the analysis, no translations of these are provided.	Russian translations of lemmas are provided as a part of the analysis.
Words are tagged in the analysis, but not segmented.	The analysis indicates morpheme boundaries.
For morphologically ambiguous word forms, only the form deemed the most likely by the software is provided and found by the search engine.	For morphologically ambiguous word forms, all interpretations yielded by the software are provided and found by the search engine.
Allows for diachronic analyses.	Do not allow for diachronic analyses.
Does not allow for sociolinguistic comparisons, but allows for genre comparisons (e.g., fiction vs. non-fiction).	Allow for sociolinguistic comparisons (literary language vs. social media).

Table 4: Some attributes of the infrastructures used

As an additional tool for analysing data is Vienna-based Mari Web Project's (MWP) morphological analyser found at morph.mari-language.com which yields fully-fledged interlinearisations (including all ambiguity) showing: morph realization, base morpheme (due to the concatenative nature of Mari morphology mostly, but not always, the same as the realization), gloss (including English translation), part of speech/type of suffix. Table 5 illustrates how the same forms are annotated in the tools under consideration. Highlighting indicates correct interpretations of the form at hand.

Glossed by hand	CLM	VK Corpora	MWP
(1) йолташем-влаклан *joltaš-em-βlak-lan* friend-1SG-PL-DAT 'to my friends'	part-of-speech: noun grammatical analysis: N.Pl.Dat.PxSg1.So_PNC dependency relation: HNOUN baseform: йолташ	**йолташем-влаклан** **йолташ** N anim, hum йолташ-ем-влак-лан STEM-1SG-PL-DAT gr: pl, dat, 1sg trans_ru: товарищ; друг	йолташем-влаклан йолташ -ем -влак-лан *йолташ -ем -влак-лан* *friend* -1SG -PL -DAT no -poss -num -case
(2) ачатше ača-t-še father-2sg-3sg 'as for your father'	part-of-speech: noun grammatical analysis: N.Sg.Nom.PxSg2.Foc_Poss dependency relation: HNOUN baseform: ача	**ачатше** **ача** N anim, hum ача-т-ше STEM-2SG-3SG gr: sg, nom, 2sg, 3sg trans_ru: отец; свёкор	ачатше ача -т -ше ача -ет -жЕ father -2SG -3SG no -poss -poss
(3) пӧртыштем *pört-ə̑št-em* house-INE-1SG 'in my house'	part-of-speech: noun grammatical analysis: N.Sg.Ine.PxSg1.So_CP dependency relation: ADVL→ baseform: пӧрт	**пӧртыштем** 1. **пӧрт** N пӧрт-ыште-м STEM-LOC-1SG gr: sg, loc, 1sg trans_ru: дом 2. **пӧрт** N пӧрт-ышт-ем STEM-3PL-1SG gr: sg, nom, 3pl, 1sg trans.ru: дом 3. **пӧртыш** N пӧртыш-те-м STEM-LOC-ACC gr: case.comp, sg, loc, acc trans.ru: вертячка 4. **пӧртыш** N пӧртыш-те-м STEM-LOC-1SG gr: sg, loc, 1sg trans.ru: вертячка 5. **пӧрт** N пӧрт-ыште-м STEM-LOC-ACC gr: case.comp, sg, loc, acc trans_ru: дом	пӧртыштем пӧрт -ышт -ем *пӧрт -штЕ -ем* house -INE -1SG no -case -poss пӧртыштем пӧртыш -т -ем *пӧртыш -штЕ -ем* coenurosis -INE -1SG no -case -poss

(4) шомакланда *šomak-lan-da* word-DAT-2PL 'to/for your word'	part-of-speech: noun grammatical analysis: N.Sg.Dat.PxPl2.So_CP dependency relation: X baseform: шомак	**шомакланда** 1. **шомакланаш** V шомаклан-д-а STEM-CAUS-NPST.3SG gr: npst, 3, sg, caus, caus_t trans_ru: разговаривать; ругаться 2. **шомак** N шомак-лан-да STEM-DAT-2PL gr: sg, dat, 2pl trans_ru: слово	шомакланда шомак -лан -да *шомак* *-лан* *-да* *word* -DAT -2PL no -case -poss
(5) енланат *eŋ-lan=at* person-DAT=ADD 'also to/for a person'	part-of-speech: noun grammatical analysis: N.Sg.Cmpr.PxPl1.So_CP.Foc_at dependency relation: HNOUN baseform: ең	**ең-ланат** 1. **ең** N anim, hum ең-ла-на-т STEM-SIM-1PL-ADD gr: add, sg, sim, 1pl trans_ru: человек 2. **ең** N anim, hum ең-лан-ат STEM-DAT-ADD gr: add, sg, dat trans_ru: человек	ең-ланат ең -лан -ат *ең* *-лан* *-ат* person -DAT -and ad/no -case -enc ең-ланат ең -ла -на -т *ең* *-ла* *-на* *-ат* person -COMP -1PL -and ad/no -case -poss -enc ең-ланат ең -ла -на -т *ең* *-ла* *-на* *-ат* person -PL -1PL -and ad/no -num -poss -enc ең-ланат ең -ла -н -ат *ең* *-ла* *-н* *-ат* person -PL -GEN -and

Table 5: Annotation in the tools under consideration

Example (1) represents an unambiguous case: here all three infrastructures only return one (correct) interpretation. In the case of example (2) as well, the three infrastructures agree on the (only) correct interpretation of a word – a stem marked with two possessive suffixes, one in a possessive function and one in a non-possessive function – but it should be noted that CLM glosses the non-possessive 3sg suffix distinctively, as **Foc_Poss**. In example (3), MWP returns an interpretation that is technically admissible, but unlikely to an extent that its exclusion is desirable ('in my house' vs. 'in my coenurosis' – a type of tapeworm infection), while VK Corpora return several interpretations that are not admissible (e.g., those in which the inessive suffix is purportedly followed by the accusative suffix, which is not permissible in Mari grammar), but among them the correct interpretation. In (4), CLM and MWP return only the correct interpretation, while VK Corpora returns the correct interpretation beside an inadmissible one. In example (5), however, CLM returns only an incorrect interpretation, while VK Corpora and MWP return several interpretations, including the correct one.

This illustrates how from a user[4] perspective, a cross-integration of these three resources in which the strengths of all three infrastructures are combined (CLM's sturdy morphological model, VK Corpora's handling of ambiguity, MWP's cross-integration with a Mari-English lexicon and interlinearisations) would be desirable. Optimally users could choose if they wish their searches to account for ambiguity (i.e., include all possible interpretations as in VK Corpora and MWP) or not (i.e., include only the interpretation deemed most likely by the software as in CLM). It is situationally dependent if it is desirable for the data to be

[4] Here "user" is defined as "author(s) of this paper".

disambiguated by good, but not 100% reliable, mechanisms or not: for naïve users looking at common structures, this disambiguation is desirable, but for linguists looking at rare structures which one cannot assume to be disambiguated appropriately, it is not. Such linguists might instead have to rely on regular expressions (regular expression searches are supported by both CLM and VK Corpora) to look for certain endings, but this requires greater technical competence and greater familiarity with the Mari language.

These potential problems notwithstanding, both CLM and VK Corpora afford users with the possibility to search the grammatical tags shown in Table 5, and even though CLM only provides tagging and no division into morphemes, these tags include information on the ordering of morphemes (e.g., **So_PNC** for *person suffix* > *number suffix* > *case suffix*) and can thus be utilized for the task at hand.

4. Determining the frequencies

Upon a perfunctory investigation of our results, we could determine that many incorrect analyses throughout our survey were a function of frequent ambiguities pertaining to a small number of oftentimes very widely used lexemes that could also be interpreted as a shorter lexeme with a possessive suffix: *una-βlak* guest-PL 'guests' was misinterpreted as **u-na-βlak* new-1PL-PL 'our new ones', *ušem-βlak* 'unions' was misinterpreted as **uš-em-βlak* mind-1SG-PL 'my minds', *urem-βlak* 'street' was (presumably) misinterpreted as **ur-em-βlak* squirrel-1SG-PL 'my squirrels', *keremet-βlak* evil_spirit-PL 'evil spirits' was misinterpreted as **kerem-et-βlak* rope-2SG-PL 'your ropes', etc. To avert problems caused by these lexemes, we excluded them and several other ones that created ambiguity (*koman* 'layered', *sös* (element of old name for October) from all searches (e.g., for *kerem* 'rope' by adding to the search query that the baseform is not **керем**). This list is surely not exhaustive but based on the output of the unrestricted searches we can assume that they are by far the most numerous culprits of mistakes of this type.

4.1 Overall Frequencies

In this section we aim to provide raw data on the basic (absolute) frequencies of different arrangements. For this task CLM was used due to the vastly larger amounts of data.

Case suffix and possessive suffix

There is only variation for two grammatical cases here: the dative in *-lan* and the comparative in *-la*. The survey is exacerbated by the homonymity afflicting the relevant suffixes, with the comparative suffix notably being homonymous with the plural of local meaning *-la* (see Section 1). Table 6 illustrates the arrangement of the dative suffix and comparative suffix with different possessive suffixes as returned by the software. To find, for example, all purported occurrences of the dative suffix followed by the possessive suffix first person singular (i.e., the ordering CXPX), one must search for tokens where the grammatical analysis contains **Dat.PxSg1.So_CP**. In the third person singular for the ordering CXPX, one must also add results returned by **Dat** and **Foc_Poss** as not to miss possessive suffixes used in a non-possessive function. In many cases, however, the false positive results – i.e., results returned that do not actually show the desired structure – greatly outnumber the legitimate results, as can be quickly determined by perusing the outputs and especially by viewing the most commonly found forms. These fields are highlighted in grey. It must also be noted that false negatives – i.e., tokens that should fall into these categories but are erroneously classified as something else – are by default excluded in this overview.

	Dative		Comparative	
	PXCX	**CXPX**	**PXCX**	**CXPX**
1SG	11,409	460	1,650	14,292
2SG	8,396	485	1,318	265
3SG	46,270	4,643	5,148	2,353
1PL	1,518	1,124	77	17,030
2PL	331	2,157	304	160
3PL	7,880	403	456	3,624

Table 6: Dative and Comparative (singular), PXCX vs. CXPX

Due to the high number of erroneous results, in spite of the exclusion of problematic stems, care must be applied before drawing conclusions. While the raw data shows PXCX as more common than CXPX in dative with 1PL, the critical mass of results are misinterpreted (e.g., the systematic erroneous analysis of inflected feminine patronyms such as *Petrovna-lan* Petrovna-DAT 'to [given name] Petrovna' as **Petrov-na-lan* Petrov-1PL-DAT 'to our Petrov'), with only comparatively few correctly analysed forms (e.g., *el-na-lan* country-1PL-DAT 'to/for our country') – one can thus safely assume that, in the first person plural, CXPX outnumbers PXCX here in practice. For the comparative, erroneous results greatly outnumber legitimate results regardless of the search pattern (e.g. *salam* 'greeting; hello' erroneously interpreted as **sa-la-m* scythe-CMPR-1SG 'like my scythe'), but legitimate occurrences of the arrangement PXCX can be encountered, especially if one restricts oneself to animate nouns and especially kinship terms, e.g., *aβa-m-la* mother-1SG-CMPR 'like my mother', *kokaj-na-la* aunt-1PL-CMPR 'like our aunt', and as expected CXPX can be encountered in the third person singular when the suffix is used in a non-possessive function, e.g. *ruš-la-že* Russian-CMPR-3SG 'in Russian, on the other hand'.

Despite the uncertainties encountered here, the following conclusions can be drawn:

- In the dative, the arrangement PXCX is dominant in all persons but 1PL and 2PL, where CXPX is dominant. The greatest variation can be found in 3SG, where PXCX dominates greatly but an ample body of CXPX examples can be found. Here the alternation can be assumed to be determined by the function of the possessive suffix: when used possessively, PXCX is dominant, while non-possessively used suffixes occur in the arrangement CXPX.
- An explanation is required for the unusual, but not rare, deviations from the dominant suffix orderings as represented by forms such as *el-na-lan* country-1PL-DAT 'to/for our country'.
- Legitimate usage examples of comparative suffix with possessive suffixes seems to be rare, but it is easier to find examples of the arrangement PXCX than CXPX.

Number suffix and possessive suffix

As the associative plural in *-mə̂t* and the plural of local meaning in *-la* are not subject to any notable variation (see Section 1), our investigation here is restricted to the plural suffixes *-βlak* and *-šamə̂ć*. To find all cases of the possessive suffix first person singular co-occurring with a plural suffix but no case suffix (i.e., nouns in the nominative) and ordering PXNX, one must search for all tokens with a grammatical analysis containing **Nom.PxSg1.So_NP**. If one wishes to restrict oneself to one suffix or the other, one must furthermore search for tokens containing their Cyrillic realizations, **влак** or **шамыч**, as the tagging does not distinguish between them. As these suffixes are not subject to any allomorphy, this is not problematic. In the third person singular for NXPX, one must again add results returned by **влак** / **шамыч** co-occurring with **Foc_Poss** in the grammatical analysis to include non-possessive usages of this possessive suffix.

In general, this point of investigation is afflicted by considerably less ambiguity, given that the plural suffixes *-βlak* and *-šamə̂ć* are not subject to homonymity and given that they are orthographically preceded by a hyphen. Table 7 shows the frequency of the different arrangements for all possessive suffixes and both plural suffixes, with problematic stems excluded.

	-βlak		*-šamə̑ć́*	
	PXNX	**NXPX**	**PXNX**	**NXPX**
1SG	2,924	473	611	158
2SG	1,452	558	188	279
3SG	14,835	1,496	1,087	465
1PL	2,979	282	215	86
2PL	376	69	39	21
3PL	444	164	27	50

Table 7: Number suffixes and possessive suffixes, PXNX vs. NXPX, with problematic stems excluded

For both *-βlak* and *-šamə̑ć́*, the ordering PXNX is considerably more common than NXPX – with the exception of 2SG and 3PL in the case of *-šamə̑ć́*, where the ordering NXPX is considerably more common. A perfunctory analysis of the results suggests a possible explanation for 2SG: numerous results show the possessive suffix used in a non-possessive function, and non-possessive usage of these suffixes is known to coincide with the final placement of the suffix. It would thus suggest itself that the non-possessive usage of 2SG is typical of the very same dialects for which the suffix *-šamə̑ć́* is typical. We have no good explanation for NXPX in 3PL at this point.

Number suffix, Possessive suffix, case suffix

Even though we, based on previous research, had made some a priori assumptions on the inadmissibility of certain suffix arrangements, for the sake of completeness, here we will investigate all theoretically possible permutations of possessive suffixes, number suffixes, and case suffixes. Given the sparsity of data even in an exceedingly large corpus when looking at unusual combinations, we have not distinguished between the plurals in *-βlak* and *-šamə̑ć́* here. To find all examples of the genitive suffix followed by the plural suffix followed by the possessive suffix 1SG (i.e., the ordering CXPXNX), one must search for grammatical analyses containing **Gen.PxSg1.So_CPN**. As above, problematic stems are systematically excluded from the search queries. Given the small size of the output, we could manually investigate the results and separate appropriate findings from erroneous ones; Table 8 only shows the findings we considered correctly identified.

Once again, the comparative with its ending *-la* was the most problematic, specifically when looking at 1PL with the arrangement NXCXPX: the overwhelming mass of the 1,010 results returned by this query were false analyses of dative forms with the additive clitic *=at*, e.g., *pašajeŋ-βlak-lan=at* worker-PL-DAT=ADD ‘also to/for the workers’ was erroneously analysed as *pašajeŋ-βlak-la-na=t* worker-PL-CMPR-2PL=ADD ‘also like our workers’. To exclude all false analyses of this type (i.e., all word forms ending in *-lanat*), we had to utilize regular expressions, as there is not currently a “does not include” functionality in the search mask. Such an addition would be desirable. After the exclusion of these forms, none of the remaining forms were plausible comparative forms.

Only for the genitive, accusative, and dative cases could we find enough data to allow for a meaningful analysis. In all cases PXNXCX and NXPXCX are admissible, with the first variant dominating over the second. There seem to be significant differences in the ratio between the two arrangements depending on the person, but we do not at this point dare to assume if this is noise in the data or if these differences are significant and due to a functional explanation currently eluding us (e.g., the comparatively high share of NXPXCX forms in 2SG: does this alternation relate to the non-possessive usage of this suffix even when the suffix is not used in the final position in either arrangement?)

In the dative, examples of the arrangement NXCXPX can be found as well, especially in 3SG. Here it seems likely that the usage of this arrangement relates to the non-possessive usage of this suffix.

		PXNXCX	PXCXNX	NXCXPX	NXPXCX	CXPXNX	CXNXPX
GENITIVE	1SG	196	0	0	21	0	0
	2SG	46	0	0	22	0	0
	3SG	1,421	0	0	38	0	0
	1PL	725	0	0	67	0	0
	2PL	39	0	0	18	0	0
	3PL	40	0	0	13	0	0
ACCUSATIVE	1SG	517	0	0	60	0	0
	2SG	328	0	0	224	0	0
	3SG	4,024	0	0	643	0	0
	1PL	485	0	0	96	0	0
	2PL	106	0	0	31	0	0
	3PL	125	0	0	55	0	0
COMITATIVE	1SG	0	0	0	0	0	0
	2SG	0	0	0	0	0	0
	3SG	3	0	0	2	0	0
	1PL	0	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	0	0	0	0	0	0
DATIVE	1SG	334	0	1	34	0	0
	2SG	148	0	3	37	0	0
	3SG	1,614	0	32	28	0	0
	1PL	252	0	5	1	0	0
	2PL	41	0	8	0	0	0
	3PL	68	0	0	8	0	0
COMPAR.	1SG	4	0	0	1	0	0
	2SG	0	0	0	0	0	0
	3SG	2	0	0	0	0	0
	1PL	1	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	0	0	0	0	0	0
INESSIVE	1SG	0	0	4	0	0	0
	2SG	0	0	0	0	0	0
	3SG	2	0	7	0	0	0
	1PL	0	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	0	0	0	0	0	0
ILLATIVE	1SG	0	0	0	0	0	0
	2SG	0	0	0	0	0	0
	3SG	2	0	0	0	0	0
	1PL	0	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	1	0	0	0	0	0
LATIVE	1SG	0	0	0	0	0	0
	2SG	0	0	0	0	0	0
	3SG	2	0	0	0	0	0
	1PL	0	0	0	0	0	0
	2PL	0	0	0	0	0	0
	3PL	0	0	0	0	0	0

Table 8: All arrangements of PX, NX, and CX for all non-nominative cases

4.2 Diachronic Comparison

Our next point of investigation is changes in the observed frequencies over the course of time. Only CLM with its large time depth affords the means to do this in a meaningful way. For the pilot study at hand, we restricted ourselves at comparing texts from before and after the year 2000. Given the sparsity of unambiguous forms with many arrangements, we are restricting ourselves to comparisons for which we can find sufficient data to render a comparison meaningful. We also did not differentiate between the plural suffixes *-βlak* and *-šamə̑č* here.

Case suffix and possessive suffix

	Before 2000		After 2000	
	PXCX	CXPX	PXCX	CXPX
1SG	4,158	125	4,925	178
2SG	3,015	170	3,328	190
3SG	13,368	429	18,875	664
1PL	101	259	97	392
2PL	8	394	62	256
3PL	479	9	2,770	11

Table 9: PXCX ~ CXPX in the dative, before and after 2000

Number suffix and possessive suffix

	Before 2000		After 2000	
	PXNX	NXPX	PXNX	NXPX
1SG	1,324	287	1,944	269
2SG	611	298	882	346
3SG	4,257	523	9,913	669
1PL	1,694	230	3,115	102
2PL	93	24	280	53
3PL	216	97	204	83

Table 10: PXNX ~ NXPX, before and after 2000

Number suffix, Possessive suffix, case suffix

		Before 2000		After 2000	
		PXNXCX	NXPXCX	PXNXCX	NXPXCX
GENITIVE	1SG	81	14	215	5
	2SG	163	8	202	9
	3SG	305	14	950	15
	1PL	362	39	470	20
	2PL	8	14	25	3
	3PL	11	7	25	5
ACCUSATIVE	1SG	70	8	408	36
	2SG	212	70	302	107
	3SG	1,194	226	2,351	302
	1PL	455	53	1,051	31
	2PL	32	17	69	10
	3PL	55	20	54	21
DATIVE	1SG	114	13	233	15
	2SG	72	9	253	20
	3SG	451	10	985	12
	1PL	175	1	424	0
	2PL	11	0	27	0
	3PL	19	5	39	2

Table 11: All arrangements of PX, NX, and CX (excerpt), before and after 2000

The data indicates a decrease in variation over the passage of time, with PXNX having an 85% share before 2000, but a 91% share after 2000. We are not confident that other trends that can be observed are particularly noteworthy, given the small amount of data.

4.3 Literary language vs. social media

In this section we will compare forms found in VK-MAIN and VK-SMC to compare variation as found in literary Meadow Mari with variation found in Mari as it is used on social media platforms. We restricted ourselves to analysing word forms not considered ambiguous by this infrastructure, which without a doubt excluded a lot of otherwise useful data, esp. considering the false analyses oftentimes included in these infrastructures as illustrated in Section 2. Here again we have not distinguished between the plural suffixes *-βlak* and *-šamâč* due to the relative paucity of data.

Case suffix and possessive suffix

	Literary		Social Media	
	PXCX	**CXPX**	**PXCX**	**CXPX**
1SG	843	0	1,014	3
2SG	142	2	874	6
3SG	4,942	0	1,970	0
1PL	42	159	532	182
2PL	8	82	90	426
3PL	1,045	2	496	1

Table 12: PXCX ~ CXPX in the dative, literary and social media texts

Number suffix and possessive suffix

	Literary		Social Media	
	PXNX	**NXPX**	**PXNX**	**NXPX**
1SG	508	39	269	40
2SG	59	25	37	49
3SG	6,328	144	1,066	129
1PL	996	12	335	87
2PL	86	1	53	30
3PL	147	19	58	24

Table 13: PXNX ~ NXPX, literary and social media texts

Number suffix, Possessive suffix, case suffix

		Literary		Social Media	
		PXNXCX	**NXPXCX**	**PXNXCX**	**NXPXCX**
GENITIVE	**1SG**	53	2	4	0
	2SG	3	3	1	0
	3SG	553	4	69	3
	1PL	240	4	55	4
	2PL	12	1	2	0
	3PL	7	9	3	0
ACCUSATIVE	**1SG**	149	4	16	3
	2SG	18	8	10	18
	3SG	1,213	71	167	45
	1PL	132	9	63	20
	2PL	38	0	42	20
	3PL	25	10	19	5
DATIVE	**1SG**	91	3	27	2
	2SG	12	1	7	4
	3SG	609	2	56	2
	1PL	83	0	40	0
	2PL	22	0	14	0
	3PL	20	1	4	3

Table 14: All arrangements of PX, NX, and CX (excerpt), literary and social media texts

For the most part, the same arrangements are predominant in both genres, but the dominance is weaker in social media texts – that is to say, there is more variation in these. In the case of the dative, PXCX has a 97% dominance in VK-MAIN but only 89% in VK-SMC. 2SG is especially notable here: here the arrangement CXPX becomes dominant. The non-possessive usage of this suffix might serve as an explanation here, considering that its usage seems to be considered dialectal and colloquial. In sharp contrast to this, in the case of 1PL, the arrangement PXCX is dominant in social media (75%), while in literary texts CXPX dominates (79%). This might be interpreted as a case of paradigmatic levelling (with the deviation of 1PL and 2PL from other persons previously noted seemingly disappearing in colloquial speech), but curiously the same phenomenon cannot be observed in 2PL. Further investigation is necessary here.

When looking at combinations of three suffixes, an increase in variation can be observed as well. For the accusative, PXNXCX dominates with 94% in VK-MAIN, but only 74% in VK-SMC.

5. Outlook and conclusions

The results yielded by our corpus-based survey line up with the statistics assembled by Jorma Luutonen over 20 years and published in his 1997 dissertation.

As regards our desire for functional explanations of this variation, the usage of possessive suffixes 2SG and 3SG not only as possessive markers but also as determining elements of sorts serves as a plausible explanation for some of the variation encountered. It does not, however, serve as an explanation for variation encountered in relation to other suffixes, nor can it explain the three-way distinction PXNXCX ~ NXCXPX ~ NXPXCX.

We cannot honestly claim to have made serious headway into explaining this variation, but the binary question raised at the outset of this study – if novel electronic tools such as the corpora under investigation here can serve as tools of analysis in this domain – can be answered with a resounding "yes". Along the way, however, we encountered shortcomings in the infrastructures we were using that are quite systematic,

and thus can be assumed to be quite straight-forward to fix: if forms are analysed incorrectly as a rule, one need only change that rule to improve the output.

It would be desirable to revisit the data contained in the corpora with some concrete hypotheses as regards the alternation encountered that can be verified or falsified based on the data. While we do not at this point have these, there are starting points for future investigations we are considering.

Does the sound structure of the base word and suffix matter?

Could phonological or prosodic elements be influencing the choice of one version or another? That the possessive suffixes 1PL (*-na*) and 2PL (*-da*) deviate from all other possessive suffixes seems notable in this respect as these two are phonologically very similar to one another, but different from others: they have an onset consonant followed by the vowel *a*, and always form a distinct generally stressed syllable of their own when attached to a stem. When investigating the variation that can be found, it might be worth investigating if systematic differences can be observed based on the sound structure of the stems taking suffixes.

Does the function of the dative matter?

The dative case has a wide range of meanings in Mari (cf. Alhoniemi 1985: 52–54): it can mark the indirect object of an action ('I gave a book to my friend'), a benefactor ('I baked a cake for my squirrel[5]'), or a purpose ('I went to the well for water'). It can be directly governed by the argument structure of a syntactically superordinate verb (e.g., 'to help', 'to call (by phone)'). Does the function of the dative have an influence on the choice of suffix arrangement? This seems especially worthy of investigation given the unusually large variation encountered in the dative case.

Does the sentence structure matter?

Syntactic rules in Mari can be observed to lose their rigidity over distance within an example – i.e., they are subject to saliency effects. The following example sentence, taken from a Mari textbook, seems to violate a Mari grammatical rule according to which quantifiers such as *βiċ́* 'five', *šuko* 'many', or *ikmə̑ńar* 'a few' co-occur with singular rather than plural forms (cf. Riese et al. 2019: 56).

(2) Meadow Mari (Riese et al. 2017: 168)

ikmə̑ńar	joltaš-em,	poškud-em	da	rodo-βlak-em
a_few	friend-PXSG1	neighbour-PXSG1	and	relative-PL-PXSG1

'a few of my friends, neighbours and relatives'

Consultations with native speakers confirmed that this is not a typo; the plural marking of *rodo* 'relative' is admissible here in spite of its co-occurrence with *ikmə̑ńar* 'a few'. However, our native informants rejected **ikmə̑ńar rodo-βlak* as completely ungrammatical: only the distance between the quantifier and the quantified, i.e., the quantifier no longer being salient when the quantified is verbalized, makes the usage of a plural suffix admissible. Likewise, the salience of the possessor might influence the manner in which possessive suffixes are verbalized. Future investigations could investigate suffix ordering in relation with the distance between possessor and possessum in a clause.

[5] One reviewer objected to this example sentence as a squirrel seemed like a semantically unlikely benefactor in this example. However, both authors of this paper have at some point shared their lives with a squirrel. As part of the revision process, one of the authors, to ensure the naturalness of this example sentence, baked a cake for their squirrel.

Glossing abbreviations

1	1st person	DAT	dative
2	2nd person	GEN	genitive
3	3rd person	INE	inessive
ACC	accusative	LOC	local
ADD	additive	NX	number suffix
ASS	associative	PL	plural
CMPR	comparative case	PX	possessive suffix
COMP	comparative degree	SG	singular
CX	case suffix		

References

Alhoniemi, Alho. 1985. *Marin kielioppi*. Apuneuvoja suomalais-ugrilaisten kielten opintoja varten 10. Helsinki: Suomalais-Ugrilainen Seura. ISBN: 951901988X

Arkhangelskiy, Timofey. 2019. Corpora of social media in minority Uralic languages. *Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages*, 125–140. Tartu. https://doi.org/10.18653/v1/W19-0311

Caballero, Gabriela. 2010. Scope, Phonology and Morphology in an Agglutinating Language: Choguita Raràmuri (Tarahumara) Variable Suffix Ordering. *Morphology* 20(1): 165–204. https://doi.org/10.1007/s11525-010-9147-4

Guseva, Elina and Philipp Weisser. 2018. Postsyntactic reordering in the Mari nominal domain: Evidence from Suspended Affixation. *Natural Language & Linguistic Theory* 36(4): 1089–1127. https://doi.org/10.1007/s11049-018-9403-6

Honti, László. 1995. Zur Morphotaktik und Morphosyntax der uralischen/finnisch-ugrischen Grundsprache. In *Congressus Octavus lnternationalis Fenno-Ugristarum 10.-15.8.1995. Pars I. Orationes plenariae et conspectus quinquennales*, edited by Heikki Leskinen, pp. 53–82. Gummerus, Jyväskylä. ISBN: 9529066848

Luutonen, Jorma. 1997. *The variation of morpheme order in Mari declension.* Mémoires de La Société Finno-Ougrienne 226. Helsinki: Suomalais-ugrilainen seura. ISBN: 9525150011

Rice, Keren. 2000. *Morpheme order and semantic scope: word formation in the Athapaskan verb* (Cambridge studies in linguistics 90). Cambridge University Press, Cambridge/New York. ISBN: 9780521024501 https://doi.org/10.1017/CBO9780511663659

Riese, Timothy, Jeremy Bradley, Monika Schötschel, and Tatiana Yefremova. 2019. *Mari (марий йылме): An Essential Grammar for International Learners. [Draft version]*. University of Vienna. grammar.mari-language.com

Riese, Timothy, Jeremy Bradley, Emma Yakimova & Galina Krylova. 2017. *Оҥай марий йылме: A Comprehensive Introduction to the Mari Language.* 3.2. Vienna: Department of Finno-Ugrian Studies, University of Vienna. omj.mari-language.com

Simonenko, Alexandra. 2014. Microvariation in Finno-Ugric possessive markers. In *Proceedings of the 43rd annual meeting of the North East Linguistic Society*, edited by Hsin-Lun Huang, Ethan Poole, and Amanda Rysling, pp. 127–140. Graduate Linguistics Student Association, Amherst, MA. ISBN: 149751066X

Tauli, Valter. 1953. The sequence of the possessive suffix and the case suffix in the Uralian languages. *Orbis* II:2. 397–404.

Trosterud, Trond and Valerij Alikov. 1994. Марийское двуязычие у мари. In *Volgalaiskielet muutoksessa: volgalaiskielten symposiumi, Turussa 1.–2.9.1993*, edited by Arto Moisio and Jaana Magnusson, pp. 111–117. Turun Yliopiston Suomalaisen ja Yleisen Kielitieteen Laitoksen julkaisuja 45. University of Turku. ISBN: 9512901765

Den historiske utviklinga til preaspirasjon i samiske språk

Pavel Iosad
Universitetet i Edinburgh

Abstract

Preaspiration of voiceless stops is a well-known feature of the phonological systems of the Sámi languages, as in Northern Sámi [jahki] *jahki* 'year'. In some form, it is found in all Sámi varieties that we have records of, with the sole exception of Inari Sámi. We also possess a good understanding of how Sámi preaspiration is related to the consonant gradation systems of other Uralic languages, in particular those of the Finnic varieties. Present-day Sámi languages differ somewhat in the role that preaspirated stops play in their phonological systems. In this paper I outline how this variation arose in the course of the historical development from Proto-Sámi to the present-day languages. I propose a scenario grounded in the *life cycle of phonological processes* (Kiparsky 1995, Bermúdez-Otero 2007, 2015). I argue that this framework is especially well suited to clarifying diverse developments of phonetic and phonological patterns from a common historical source, and is applicable also to Sámi material. I justify this approach by tracing the development of preaspiration from a phonetic rule of Proto-Sámi to the various outcomes attested in the present-day languages.

Keywords: phonology, sámi, historical linguistics

1. Preaspirasjon i dei moderne samiske språka

Preaspirasjon vert vanlegvis definert som ein periode med ustemd friksjon framfor ein konsonant, oftast ein stemmelaus klusil. For det meste finn vi preaspirasjon framfor ustemde klusilar, men også frikative konsonantar kan vere preaspirerte. Det «prototypiske» tilfellet er at friksjonen vert artikulert som ein glottal lyd av [h]-typen, kan hende med ein del kontekstuell innverknad frå nabolydane, iallfall etter ein vokal.

Det er godt kjent at preaspirerte klusilar spelar ein viktig rolle i det fonologiske og morfologiske systemet til dei fleste samiske språka. Preaspirasjon kan stå for både leksikalske og grammatiske forskjellar: SaaN [ˈtik:i] *dikki* 'ting.GEN' ≠ [ˈtihki] *dihki* 'lus.GEN' er eit leksikalsk minimalt par, medan i [ˈn͡eahpi] *neahpi* 'onkelbarn.NOM' ≠ [ˈneapi] *neabi* 'onkelbarn.ACC' ser vi bruken av preaspirasjon som grammatisk uttrykksmiddel.

Det vert ofte hevda at preaspirasjon er eit sjeldant fenomen blant språka i verda (sjå t.d. Wagner 1964, Silverman 2003, Blevins 2017). I denne konteksten er det spesielt merkverdig at preaspirasjon er så utbreidd på Nordkalotten, altså både i samisk og nordisk (Gunnar Ólafur Hansson 2001, Pétur Helgason 2002), i tillegg til keltiske språk som skotsk-gælisk. Preaspirasjon vart dimed tidleg noko som fleire fagfolk peika på som eit mogleg utfall av språkkontakt i regionen (t.d. Posti 1954, Wagner 1964, Kylstra 1967, 1972). Om vi skal evaluere desse påstandane, er det viktig å ta grep om spørsmålet kor preaspirasjonen kjem frå. Mykje har vore sagt om opphavet til preaspirasjon i nordisk — ikkje minst av dei som har hevda at dette draget kom inn frå eit samisk substrat (Rießler 2004, 2008, Kusmenko 2008, Bull 2011) — men på den samiske sida er spørsmålet mindre utforska. Eg ønskjer hermed å gjere eit bidrag på dette feltet.

I dette avsnittet gir eg den nødvendige bakgrunnen om stadieveksling i både austersjøfinske og samiske språk (avsnitt 1.1 og avsnitt 1.2) og går deretter gjennom utviklingane som vi finn i dagens samiske varietetar (avsnitt 1.3). Avsnitt 2 skildrar det teoretiske grunnlaget for studien, og i avsnitt 3 nyttar eg teorien for å legge fram eit scenario for den historiske utviklinga av preaspirasjon i samiske språk. Til slutt kjem eg tilbake til spørsmålet om språkkontakt i avsnitt 4.

1.1. Stadieveksling og opphavet til preaspirasjon

For å forstå både kor preaspirasjon kjem frå og korleis den fungerer i dagens språksystem, hjelper det å byrje med å avklare mønstera for *stadieveksling* i uralske (spesielt samiske og austersjøfinske) språk. Det er

©2022 Pavel Iosad. *Nordlyd* 46.1: 75-101, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigert av Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publisert ved UiT Noregs arktiske universitet.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6350

enklast å byrje med stadieveksling i postvokaliske klusilar.

I det uralske urspråket (jf. Sammallahti 1988, Luobbal Sámmol Sámmol Ánte 2022) fanst det berre ein serie klusilar, som var stemmelause i framlyd.[1] I innlyd kunne klusilane **p t k* (og affrikatane **c č*, som vi ikkje tek med her, men som i grunnen fungerer på same måte) vere både korte og lange: PUr **joke* 'elv' (C-rekkja), **appe* 'svigerfar' (CC-rekkja).

Stadieveksling er eit sams drag for austersjøfinske og samiske språk. Her fokuserer vi for det meste på den såkalla *stavingsvekslinga* eller *rotvekslinga*. Den går ut på at ein konsonant i visse stillingar i innlyd, anten den var kort eller lang, kunne stå i *sterk* eller *svak* grad, eller stadium, alt avhengig av strukturen til den neste stavinga. Innlydskonsonanten var sterk om den neste stavinga var open (altså slutta på vokal), noko som tradisjonelt vert notert med teiknet <ˋ> over konsonanten, og svak om stavinga var lukka; den svake graden vert notert med <ˇ>. I dei to orda ovanfor finn vi det sterke stadiet i NOM.SG (**jokè*, **ap̀pe*) og det svake stadiet til dømes i GEN.SG, der suffikset **-n* gjer stavinga til lukka (**joǩen*, **ap̌pen*). Dimed har vi fire moglege utfall av ein ururalsk klusil: svak kort (Č), sterk kort (Ċ), svak lang (ČC) og sterk lang (ĊC). Historiske lange klusilar skal vi også kunne kalle *geminatar*.

Dette firedelte opphavlege systemet viser for det meste to- eller tredelte utfall i dei fleste moderne språka. Til dømes i det finske standardspråket finn vi vekslingsmønstera som svarar til den sterke/svake alternasjonen i enkeltklusilrekkja (som i *joki* 'elv' ∼ *joet* 'elv.PL') eller til den same alternasjonen i geminatrekkja (som i *kukka* 'blomster' ∼ *kukat* 'blomster.PL'). Dette er fordi alternasjonar mellom korte og lange klusilar mangla i urspråket, og oppstod aldri i språk som standardfinsk. Slike alternasjonar kunne derimot oppstå som følgje av nokre seinare prosessar, blant anna lenging framfor lange vokalar, som vi ikkje skal gå inn på her. I neste avsnitt ser vi på korleis dette systemet fungerer i samiske språk.

1.2. Stadieveksling og preaspirasjon i samiske språk

I dei fleste samiske språka fell det sterke stadiet av enkeltklusilar saman med det svake stadiet av geminatar. Ord som SaaN *johka* 'elv.NOM.SG' representerer gamle korte klusilar: i sterkt stadium finn vi i mange språk *korte* preaspirerte klusilar, som i dagens standardortografi vert skrivne med <hp ht hk>, til motsetnad til det svake stadiet, som i SaaN *joga* 'elv.GEN.SG', der vi finn ikkje-preaspirerte klusilar eller — liksom i mange austersjøfinske språk — diverse typar lyd som kan vere vidare utviklingar av desse (til dømes vert SaaN *joga* uttala med [k], [g], [v] osv, alt etter dialekt og iblant talar). I ord som SaaN *vuohppa* 'svigerfar.NOM.SG' ser vi utfall av gamle lange klusilar. I sterkt stadium står det — i nordsamisk — *lange* preaspirerte klusilar, som vert skrivne <hpp htt hkk>. Ortografien speglar her den tradisjonelle skildringa, som hevdar at det er lengda på lukkingsfasen som er skilnaden mellom lange og korte preaspirerte klusilar, men som vi skal sjå nedanfor er biletet meir innfløkt. I alle høve er stoda slik at denne lange preaspirerte klusilen i ord som *vuohppa* vekslar i svakt stadium med den same typen kort preaspirert klusil som vi finn i det sterke stadiet av ord som *johka*, altså SaaN *vuohpa* 'svigerfar.GEN.SG'.

Tradisjonelt seier vi at ord som *joga* representerer kvantitetsgraden Q1 («kort»), ord som *johka* og *vuohpa* er Q2 («lang»), og ord som *vuohppa* er Q3 («overlang»); dimed finn vi vekslingsmønster Q1–Q2 i 'elv' og Q2–Q3 i 'svigerfar'. Eit tredje mønster er Q1–Q3. Historisk oppstår dette når ein opphavleg enkeltkonsonant i sterkt stadium (som vi ventar skal få Q2, og som vekslar regelrett med Q1) vart lengd til geminat, som i sterkt stadium gir Q3. Denne lenginga skjer for det meste framfor ymse typar lange vokalar, liksom i fleire austersjøfinske språk, t. d. i SaaN *bohcco* 'reindyr.GEN' frå **pōcōj-i-n* (svakt stadium *boazu* 'reindyr.NOM' < **pōcōj*). I tillegg har fleire nordsamiske dialektar (sjå t. d. om Eanodat Sammallahti 1977, om Guovdageaidnu Bals Baal, Odden & Rice 2012) andre typar lenging til Q3.

Til forskjell frå dei austersjøfinske språka er stoda i samisk slik at det ikkje er berre klusilar og affrikatar som tek del i stadieveksling. Både frikativar som [s] og sonorantar som [l] eller [r] kan vere både korte, lange og overlange. I tillegg finn vi ein skilnad mellom sterke og svake stadium av både lange og korte konsonantar, spesielt nasalar, som er heilt parallell den vi har ved klusilar, med samanfall av dei to «midtre» gruppene: jf.

[1] Her ser vi bort frå konsonanten **δ*, som kan ha vore [d] i det minste i nokre tilfelle.

		*p t k		*pp tt kk	
Kontekst		p̆ t̆ k̆	p̀ t̀ k̀	p̆p t̆t k̆k	p̀p t̀t k̀k
Etter vokal	Utfall	b d g	b͡p d͡t g͡k	(b)p (d)t (g)k	hp ht hk
	IPA	jogɐ	jog͡k	tɨg͡kʲɛ	tɨhkʲɛ
	Original	*jogą*	*joɢkᴬ*	*tį̇ǫ̇kɛ*	*tį̇ɔ̄k̨ɛ*
	Glose	‘elv.GEN’	‘elv’	‘lus.GEN’	‘lus’
Etter sonorant	Utfall	b d g	b͡p d͡t g͡k	(b)p: (d)t: (g)k:	hp ht hk
	IPA	ŋaːrgˑɐ	ŋaːrːg͡k	toːrkɐ	toɽːk
	Original	*ńārɔ̀ą*	*ńārᵍ̄k(ᴬ)*	*tōrᵍk̄ą*	*torˑ̄k(ᴬ)*
	Glose	‘nes.GEN’	‘nes’	‘pelskåpe.GEN’	‘pelskåpe’

Tabell 1: Stadieveksling og klusilar i tersamisk etter T. I. Itkonen (2011)

SaaN *biebmu* ‘mat.NOM’, svakt stadium *biepmu* ‘mat.ACC’ (PSaa **pēm̀mō*, **pēm̆mōm*); sterkt stadium *liepma* ‘buljong.NOM’, svakt stadium *liema* ‘buljong.ACC’ (PSaa **lēm̀ę*, **lēm̆ęm*).

Det er også viktig å peike på dei litt annleis høva i ikkje-postvokalisk stilling. Etter ei lukka staving, spesielt etter ein sonorant, fell dei to rekkjene C̀ og C̆C *ikkje* saman: jf. SaaN *gánnda* ‘gutt.NOM’ (sterkt stadium av nasal + enkeltlyd) men *gumppe* ‘ulv.GEN’ (svakt stadium av nasal + geminat). Forskjellen mellom rekkjene og gradane kjem til syne på ymse måtar som famnar både lengd, preaspirasjon, nærvær av stemme eller stemmeløyse, og innskotsvokal, alt etter språk og type av konsonantklynge.

1.3. Utviklingar i samiske språk

I dette avsnittet skal vi sjå nærmare på korleis dette grunnleggjande systemet fungerer i dei særskilde samiske språka. Vi byrjar med kolasamisk og går vidare vest- og sørover. Skildringa her byggjer på oversiktsverk av Korhonen (1981) og Sammallahti (1998) og den etymologiske ordboka til Lehtiranta (1989), i tillegg til dei kjeldene som er oppgjevne for kvart einskilt språk. Oversiktstabellane viser, der kjeldegrunnlaget tillèt, utfallet av PSaa **jokę* ‘elv’ (C-rekkje etter kort vokal), **tikkē* ‘lus’ (CC-rekkje etter kort vokal), **kētę* ‘hand’ (C-rekkje etter lang vokal, der utfallet er forskjellig frå det etter kort vokal), og **ākkō* ‘bestemor’ (CC-rekkje etter lang vokal); utvalet av ord med konsonantklynger av typen «sonorant + klusil» er noko meir tilfeldig.

1.3.1. Kolasamisk

Med tanke på korleis preaspirerte klusilar utviklar seg kan vi samle dei to kolasamiske språka tersamisk og kildinsamisk under eitt. Spørsmålet om stadieveksling i desse varietetane er grundig drøfta av T. I. Itkonen (1916) (sjå også ordboka til T. I. Itkonen 2011), i tillegg til skildringane av dei særskilte språka som Kert (1971) og Rießler (2022) for kildinsamisk og Tereshkin (2002) for tersamisk.[2] Mønstera som kjem fram i T. I. Itkonen sitt verk vart også systematisk analyserte av Bańczerowski (1969).

Tabell 1 viser systemet i tersamisk som det kjem fram i ordboka til T. I. Itkonen (2011), med både dei originale transkripsjonane i det uraliske fonetiske alfabetet og mine tolkingar med det internasjonale alfabetet (IPA). Systemet er noko forenkla: T. I. Itkonen (1916) skildrar eit meir innfløkt mønster med to undertypar av Q2. Her ser vi bort frå dette, ikkje minst av di E. Itkonen (1946) si etterprøving ikkje fann dette systemet ved lag; sjå elles Iosad (under utarb.) for meir inngåande diskusjon.

Geminatar i sterkt stadium (altså Q3-graden) vert reflektert i kolasamisk som preaspirerte klusilar,

[2] Skriftspråket som er skildra hjå Kuruch (1985) skulle vere sams for alle samiske språk på den russiske sida, altså både kolasamisk i den snevre meininga (kildin- og tersamisk) og skoltesamisk (herunder akkalasamisk), som vi skildrar separat i avsnitt 1.3.2. Dette språket speglar ikkje nokon levande varietet, men byggjer for det meste på kildin-dialektane.

iallfall etter vokal. I Q1 (svakt stadium av enkeltlydar) har T. I. Itkonen (2011) stemde frikativar [β ð ɣ] som hovudvarianten, men han noterer også klusilar som moglege realisasjonar i kildinsamisk (sjå ss. xxix–xx). Russiske kjelder har som normalt ikkje-preaspirerte, ofte stemde klusilar. Dette er et forskjell frå språka vidare vestover. Som vi skal sjå, er klusilar i Q1 heller ikkje vanlege i resten av austsamisk. Det er mogleg at vi kan føre dei kolasamiske stemde klusilane (i motsetnad til uaspirerte, normalt stemmelause klusilar i Č-rekkje i vestsamiske språk) tilbake til påverknad frå russisk.

Den store forskjellen mellom kolasamisk og dei andre språka finn vi i Q2. Her fekk dei ein heller uvanleg lydtype, nemleg dei såkalla «halvstemde geminatane» <bp dt gk>. Desse vert skrivne som lange lydar som byrjar med ein stemd porsjon og sluttar med stemmelaus artikulasjon hjå t. d. T. I. Itkonen (1916, 2011), og det same finn vi i standardortografien. Kert (1971) på si side skildrar dei som lange klusilar, heilstemde utanom i utlyd. Rießler & Wilbur (2007) og Rießler (2022) skildrar dei også som stemde lange klusilar, ikkje monaleg forskjellige frå stemde geminatar i ord som SaaN *loddi* ‘fugl’.

Dette utfallet er heilt vanleg i det sterke stadiet av C-rekkja, som i SaaT [jo͡gk] ‘elv.NOM’, svakt stadium [jogɐ] ‘elv.GEN’ (<joɢkᴬ>, <jogą> [T. I. Itkonen 2011: s. 67]). Stoda er mindre klar i ČC-rekkja. Om kildinsamisk har, som resten av språkgreina, eit Q2-samanfall, ventar vi (halv)stemde geminatar også her, og slike former finst sanneleg i kjeldene, jf. SaaK [vuːhp] ‘svigerfar.NOM’, svakt stadium [vuːb͡p] ‘svigerfar.GEN’ (<vūɔ̀p̀>, <vūʙpᴬ> [T. I. Itkonen 2011: s. 790]) eller *mõhn* ‘slire.NOM’, svakt stadium *mõõn* ‘slire.GEN’ (Kuruch 1985: s. 355). Bergsland (1973: s. 66) reknar også med at kildinsamisk viser samanfallet i Q2.

Det finst også eit anna vekslingsmønster der samanfallet ikkje kjem til syne, av di ČC-konsonantane er realiserte som stemmelause, uaspirerte enkeltklusilar. Dette mønsteret er skildra som det vanlege av Kert (1971) for kildinsamisk og Tereshkin (2002) for tersamisk. Mange ord i Kuruch (1985) si ordbok viser også det same (*нāhнь* ‘trebolle’ ∼ *нāнь* ‘trebolle.GEN’). Døme av dette slaget er også å finne hjå t. d. Rießler & Wilbur (2007) og Rießler (2022). Den mest kompliserte varianten av dette systemet finn vi hjå T. I. Itkonen (1916), som vist i tabell 1 med døme frå tersamisk. Han skil C̀- og ČC-rekkjene åt kvarandre, sjølv om realiseringane deira er delvis overlappande. Førstnemnde er regelrett halvstemde geminatar, medan sistnemnde har både halvstemde og stemmelause (uaspirerte) variantar. Av denne grunnen held Bańczerowski (1969) at tersamisk, men ikkje kolasamisk, manglar Q2-samanfallet.

Det som er sams for alle desse skildringane er at ČC-rekkja etter vokal manglar preaspirasjon, til forskjell frå det sterke stadiet. Men som Sammallahti (1998) peikar på, er det heilt sannsynleg at mangelen på preaspirasjon i ČC-klusilar i kildinsamisk ikkje tyder at dei aldri hadde den. Provet for det kjem frå konsonantsambanda av typen «sonorant + geminat». I sterkt stadium vert desse reflekterte, som venta, med preaspirasjon (SaaK [peːr̥ːʰt] *nə̄ppm* ‘hus.NOM.SG’, jf. F *pirtti*, frå slavisk **рьртъ*). I svakt stadium har klusilane i T. I. Itkonen (1916) sitt system, som vist i tabell 1, dei same utfalla som i postvokalisk stilling.[3] Likevel noterer T. I. Itkonen (1916: s. xxix) at ustemde sonorantar (det vil seie preaspirerte klusilar) også er moglege i svakt stadium i desse klyngene, iallfall i kildinsamisk. Eit slikt mønster, med preaspirasjon også i svakt stadium (SaaK [per̥ʰt] *nəpm* ‘hus.GEN.SG’), er regelrett i Kuruch (1985), og vi finn tilsvarande døme hjå Kert (1971: s. 98–99), E. Itkonen (1946: s. 242), Sammallahti (1998: s. 55) og Rießler (2022).

1.3.2. Skoltesamisk og akkalasamisk

Skoltesamisk er skildra av T. I. Itkonen (1916, 2011), Korhonen, Mosnikoff & Sammallahti (1973) og Feist (2015), og McRobbie-Utasi (1991) er ein fonetisk studie over preaspirasjon i språket. I tillegg har vi Zaĭkov (1987) om akkala-dialekten i landsbyen Babinsk, som hadde fleire avvikande drag. Tabell 2 viser systemet i skoltesamisk ifølgje Korhonen, Mosnikoff & Sammallahti (1973) og Feist (2015).

I skoltesamisk finn vi i prinsipp det vanlege systemet, men det finst noko innverknad mellom kvantiteten på konsonant og vokal. Vi kan ikkje gå inn mykje nærmare på desse her (jf. E. Itkonen 1946). Éi utvikling som det er verdt å nemne er at opphavlege sekvensar «kort trykksterk vokal + kort konsonant» vert endra, med ei forlenging av konsonanten til Q3 i sterkt stadium og forlenging av vokalen i svakt stadium (dimed

[3] Mønsteret som vi ser i tabell 1, der den første konsonanten i klynga er relativt lang og den andre relativt kort i sterkt stadium, med omvendt stode i svakt stadium, er heilt typisk i samiske språk; jf. Bye (2005).

		*p t k		*pp tt kk	
Kontekst		p̆ t̆ k̆	p̀ t̀ k̀	p̆p t̆t k̆k	p̀p t̀t k̀k
Etter kort vokal (tostavingsord)	Utfall	v ð ɣɣ	(ʰp: ʰt: ʰk:)	ʰp· ʰt· ʰk·	ʰp: ʰt: ʰk:
	IPA	joˑɣˑ	joʰk:	teˑʰcˑ	teʰc:
	Ortografi	*joogg*	*jokk*	*tee´ǩǩ*	*te´ǩǩ*
	Glose	‘elv.GEN’	‘elv’	‘lus.GEN’	‘lus’
Etter lang vokal	Utfall	v ð ɣɣ	ʰp· ʰt· ʰk·		ʰp: ʰt: ʰk:
	IPA	ci͡əðˑ	ci͡əʰtˑ	laˑʰp:	laʰp:
	Ortografi	*ǩiõđđ*	*ǩiõtt*	*lääpp*	*läpp*
	Glose	‘hand.GEN’	‘hand’	‘sene.GEN’	‘sene’
Etter sonorant	Utfall	b d g	b̥ d̥ g̊	ʰp ʰt ʰk	ʰp: ʰt: ʰk:
	IPA	pe͡aldast	pĕal·d̥·an	ŋoːl̥ʰp	ŋol̥·ʰp:
	Ortografi	*peäldast*	*pe͜älddan*	njoolp	njoolpp
	Glose	‘åker.LOC’	‘åker.ESS’	‘reinsdyr.GEN’	‘reinsdyr

Tabell 2: Stadieveksling og klusilar i skoltesamisk

vekslinga *jokk* ∼ *joogg*) i tabell 2. Bortsett frå dette finn vi kortare («halvlange») preaspirerte klusilar i Q2 i skoltesamisk, som vist i tabell 2. Dei vekslar med lange preaspirerte klusilar i Q3 når dei er opphavlege geminatar, som i SaaSk [tẹhc:] *te´ǩǩ* ‘lus.NOM.SG’, svakt stadium [tẹˑʰcˑ] *tee´ǩǩ* ‘lus.GEN.SG’. I svakt stadium av opphavlege enkeltlydar finn vi i skoltesamisk stemde frikativar som [v], [ð] og [ɣ], ei typisk austsamisk utvikling; desse kan vere lange, takk vere bl. a. forlenginga som vi allereie nemnde ovanfor. Etter sonorantar har vi preaspirerte klusilar (ustemde sonorantar) i både sterkt og svakt stadium av gamle geminatar, uaspirerte (delvis stemde) klusilar frå gamle enkeltlydar.

I akkalasamisk har Zaĭkov (1987) fleire døme på uaspirerte stemmelause klusilar i ČC-rekkja, liksom i kolasamisk etter russiske kjelder, men han gir ikkje noka systematisk drøfting av fenomenet. Ein viktig forskjell mellom skoltesamisk og akkalasamisk er at sistnemnde manglar stemmelause sonorantar: preaspirasjonen vart altså missa både i sterkt og svakt stadium i konsonantklynger.

1.3.3. Enaresamisk

Dette språket (sjå t. d. Äimä 1918, E. Itkonen 1946, Bye 2007) viser det vanlege mønsteret med samanfall av sterke enkeltlydar og svake geminatar. Det spesielle med enaresamisk er at språket viser *postaspirerte* heller enn preaspirerte klusilar i både Q2 (altså resultatet av samanfallet) og Q3. Sterke geminatar er reflektert som lange, postaspirerte klusilar, medan Q2 viser enkle postaspirerte klusilar, med unnatak av den dorsale rekkja, der vi far frikativen [h] i staden for den venta [kʰ]. Dimed vekslar SaaI [tikʰːə] *tikke* ‘lus.NOM.SG’ med [tihe] *tihe* ‘lus.GEN.SG’, liksom [vu͡opʰːə] *vuoppâ* ‘svigerfar.NOM.SG’, [vu͡opʰə] *vuopâ* ‘svigerfar.GEN.SG’. I klynger har vi også postaspirasjon hjå gamle geminatar, og dimed inga stemmeløyse i sonoranten: [ˈkirkʰːo] *kirkko* ‘kirke.NOM’ ∼ [ˈkirho] *kirho* ‘kirke.GEN’. I det svake stadiet av enkeltklusilar får vi, som i skoltesamisk, stemde frikativar [v] og [ð], med [v] i staden for den venta [ɣ]: dimed [juːhə] *juhâ* ‘elv.NOM.SG’, [juːvə] *juvâ* ‘elv.GEN.SG’. Enaresamisk viser også fleire særs involverte vekselverknadar mellom kvantiteten på vokalane og konsonantane (Bye 2007), som vi ikkje går inn på her.

1.3.4. Nordsamisk

Nordsamisk er det samiske språket som vi har flest fonetiske og fonologiske studiar over; her kan vi nemne både generelle oversikt som Nielsen (1979), Nickel & Sammallahti (2011) og Luobbal Sámmol Sámmol Ánte & Ylikoski (2022), monografiske skildringar av fonologien (Sammallahti 1977, 2019) og fonetikken

		*p t k		*pp tt kk	
Kontekst		p̆ t̆ k̆	p̀ t̀ k̀	p̆p t̆t k̆k	p̀p t̀t k̀k
Etter kort vokal	Utfall	p/v ð k/ɣ	hp ht hk		h:p h:t h:k
	IPA	joka	jo:hka	ti:hki	tih:ki
	Ortografi	*joga*	*johka*	*dihki*	*dihkki*
	Glose	‘elv.GEN’	‘elv’	‘lus.GEN’	‘lus’
Etter lang vokal	Utfall	p/v ð k/ɣ	hp ht hk		h:p h:t h:k
	IPA	ki͡eða	ki͡ehta	a:hku	ah:ku
	Ortografi	*gieđa*	*giehta*	*áhku*	*áhkku*
	Glose	‘hand.GEN’	‘hand’	‘bestemor.GEN’	‘bestemor’
Etter sonorant	Utfall	p t k		hp ht hk	
	IPA	pe͡alk:i	pe͡alĕgi	pa:l̥k:a	pa:lähka
	Ortografi	*bealggi*	*bealgi*	*bálkka*	*bálká*
	Glose	‘tommel.GEN’	‘tommel’	‘betaling.GEN’	‘betaling’

Tabell 3: Stadieveksling og klusilar i nordsamisk

(Magga 1984) til einskilde dialektar, og spesialiserte studiar (Bals Baal, Odden & Rice 2006, 2012, Hiovain, Vainio & Šimko 2020).

Stadieveksling og preaspirasjon i nordsamisk byr ved første augekast på lite som er uventa. I Q3 har vi lange preaspirerte klusilar; dei er skrivne <hpp htt hkk>, men skilnaden mellom dei og dei korte klusilane <hp ht hk> som vi finn i Q2 (altså etter samanfallet av sterke enkeltklusilar og svake geminatar) ligg i lengda på preaspirasjonen, ikkje på klusilfasen; vi finn difor ei veksling mellom SaaN [tih:ki] *dihkki* ‘lus.NOM.SG’ og [tihki] *dihki* ‘lus.GEN.SG’. Dette vert stadfesta av det instrumentale studiet til Bals Baal, Odden & Rice (2006).

I det svake stadiet av enkeltklusilrekkja (Q1) har nordsamisk [ð] i koronalrekkja ([ko͡ahti] *goahti* ‘gamme.NOM.SG’ ∼ [ko͡aði] *goađi* ‘gamme.GEN.SG’). Ved andre artikulasjonsstadar finn vi anten uaspirerte klusilar (desse er ofte ustemde [p] og [k], men kan ha ein grad av klangfør artikulasjon på ein skiftande måte, alt etter talaren) eller stemde frikativar som t. d. [v], [ɣ], avhengig av dialekt: den frikative uttalen er sams for dei austlege nordsamiske dialektane og språka vidare austover. Skilnaden mellom konsonantar i Q2 og Q3, anten dei er preaspirerte klusilar eller enkeltkonsonantar som skil mellom dei to gradane, går saman med ein skilnad i lengda på den føregåande stavingskjernen: både vokalar og diftongar er korte framfor konsonantar i Q3 og lange framfor konsonantar i Q2 (Bals Baal, Odden & Rice 2006, Sammallahti 2019, Hiovain, Vainio & Šimko 2020, Luobbal Sámmol Sámmol Ánte & Ylikoski 2022); i nokre dialektar spelar kvaliteten, spesielt på diftongar, ei liknande rolle (Sammallahti 2019: s. 143–144).

Etter sonorantar er geminatar, men ikkje enkeltlydar, preaspirerte i både svakt og sterkt stadium. Fleire typar konsononatklynger med sonorant (herunder [ð]) som førsteledd har innskotsvokal i sterkt stadium, og dimed inga stemmeløyse i sonoranten sjølv som realisering av preaspirasjon. Som vanleg er andreleddet i klynga (altså klusilen, i dette tilfellet) relativt langt i svakt stadium jamført med det sterke.

1.3.5. Lulesamisk

Systemet i lulesamisk er ganske likt det nordsamiske, idet Q2, der svake geminatar fell saman med sterke enkeltlydar, vert realiserte som korte preaspirerte klusilar, og Q3 viser lange preaspirerte klusilar. I Q1 får vi uaspirerte, ustemde klusilar, dimed SaaL [jɔhkɔ] *jåhkå* ‘elv.NOM’ ∼ [jɔkɔ] *jågå* ‘elv.GEN’ og [tih:ki͡e] *dihkke* ‘lus.NOM’ ∼ [tihki͡e] *dihke* ‘lus.GEN’. Det finst få pålitelege kjelder om dei fonetiske eigenskapane til lulesamisk preaspirasjon, men ifølgje Engstrand (1987) er det slik at iallfall nokre talarar skil mellom Q2 og Q3 ved at preaspirasjonen er lengre i sistnemnde, utan nemneverdig skilnad i lengda på klusilfasen, medan andre manglar ustemd preaspirasjon i Q3. Studien kviler riktignok på ei heller beskjeden mengd data. I ei

		*p t k		*pp tt kk	
Kontekst		p̆ t̆ k̆	p̀ t̀ k̀	p̆p t̆t k̆k	p̀p t̀t k̀k
Etter kort vokal	Utfall	p t k	hp ht hk	h:p h:t h:k	
	IPA	jɔkɔ	jɔhkɔ	kah:t͡ʃaʊ	kah:t͡ʃat
	Ortografi	*jågå*	*jåhkå*	*gahttjav*	*gahttjat*
	Glose	‘elv.GEN’	‘elv	‘falle.PRS.1SG’	‘falle.INF’
Etter lang vokal	Utfall	p t k	hp ht hk		h:p h:t h:k
	IPA	ki͡etan	ki͡ehta	a:hka	a:h:ka
	Ortografi	*gieda*	*giehta*	*áhka*	*áhkka*
	Glose	‘hand.GEN’	‘hand’	‘bestemor.GEN’	‘bestemor’

Tabell 4: Stadieveksling og klusilar i pitesamisk

nyare fonetisk undersøking finn Fangel-Gustavson, Ridouane & Morén-Duolljá (2014) at alle tre kvantitetar vert haldne frå kvarandre ved hjelp av konsonantlengd i lulesamisk, men preaspirerte klusilar er ikkje med i studien deira. I konsonantklynger er situasjonen igjen ganske lik den i nordsamisk, med spesielt utbreidd bruk av innskotsvokal (jf. Larsson 1990).

1.3.6. Pitesamisk

I pitesamisk (Lagercrantz 1926, Wilbur 2014, Sjaggo 2015) finn vi igjen at det grunnleggjande systemet vert skipla av endringar i kvantitative mønster. I prinsipp har pitesamisk det same systemet som lulesamisk: uaspirerte klusilar i Q1, korte preaspirerte klusilar i Q2 og lange preaspirerte klusilar i Q3.

Det spesielle ved pitesamisk, ifølgje Sammallahti (1998: s. 21), er at vekslinga mellom Q2 og Q3 (herunder korte resp. lange preaspirert klusilar) berre er mogleg etter lange vokalar. Dette er av di geminatar vart lengde til overlange geminatar etter korte trykksterke vokalar: [ˈkah:t͡ʃat] *gahttjat* ‘falle.INF’ ∼ [ˈkah:t͡ʃaʊ] *gahttjav* ‘falle.PRS.1SG’.[4] Dette systemet kjem også fram i skildringa til Lagercrantz (1926: s. 230–231). I varietetar med dette mønsteret manglar Q2 ∼ Q3-vekslinga etter korte trykksterke vokalar generelt, også ved andre konsonantar enn klusilar og affrikatar, jf. [ˈmis:o] *misso* ‘myse’, svakt stadium [ˈmis:o] *misso* ‘myse.GEN’, jf. SaaN *mis'su* ‘myse’, svakt stadium *missu*.

Både Wilbur (2014, 2016) og Sjaggo (2015) har derimot døme på denne typen stadieveksling, som i *lihtte* ‘skål’, *lihte* ‘skål.GEN’, altså med «lulesamisk» system. Ifølgje Lehtiranta (1992: s. 33) speglar denne forskjellen ei dialektgrense, som går ved Tjidtjak. Nord for denne isoglossen vert dei sterke og svake stadia av gamle geminatar haldne åtskilde frå kvarandre etter kort trykksterk vokal, og inngår dimed i stadievekslingssystemet, medan sør for denne linja manglar vekslinga i denne konteksten. Denne forskjellen mellom korte og lange trykksterke vokalar går igjen i dei sørlege samiske språka, som vi ser på i dei neste avsnitta.

I konsonantklynger har vi det vanlege systemet med uaspirerte klusilar for historiske enkeltlydar og preaspirerte klusilar i begge stadium for gamle geminatar.

1.3.7. Umesamisk

Umesamisk fonologi viser ein god del variasjon, ikkje minst når det gjeld stadieveksling (Schlachter 1958, 1991). Larsson (2012) skildrar mange av desse på eit breitt kjeldegrunnlag, medan von Gertten (2015) gir ei oppsummering av mønstera. Samanfallet etter kort trykksterk vokal går her vidare enn i pitesamisk. Grunnen til det er at trykksterke stavingar må vere tunge, akkurat som i dei fleste norske og svenske dialektane. Når

[4] Jf. også [ˈkɛh:tot] *gähttot* ‘fortelle.INF’, der vekslinga berre inntrer når omlyd produserer ei tung trykksterk staving, sjølv om vokalen er kort i dag: [ˈkehtoʊ] *giehtov* ‘fortelle.PRS.1SG’

		*p t k		*pp tt kk	
Kontekst		p̆ t̆ k̆	p̀ t̀ k̀	p̆p t̆t k̆k	p̀p t̀t k̀k
Etter kort vokal	Utfall	h:p h:t h:k			
	IPA	juh:kən	juh:kə	tih:ki̯en	tih:ki̯e
	Ortografi	*juhkan*	*juhka*	*dihkien*	*dihkie*
	Glose	‘elv.GEN’	‘elv’	‘lus.GEN’	‘lus’
Etter lang vokal	Utfall	p t k	hp ht hk		h:p h:t h:k
	IPA	ki̯etən	ki̯ehtə	a:hka:n	a:h:ka:
	Ortografi	*giedan*	*giehta*	*áhkán*	*áhkká*
	Glose	‘hand.GEN’	‘hand’	‘bestemor.GEN’	‘bestemor’

Tabell 5: Stadieveksling og klusilar i umesamisk

ei trykksterk staving er historisk lett vert den påfølgjande konsonanten lengd, noko som minner sterkt om utviklinga i trøndsk.

Når konsonanten er ein klusil, vert utfallet av denne lenginga ein lang preaspirert klusil. Dimed finst det inga stadieveksling etter trykksterk kort vokal i ei open staving, ved anten historiske korte klusilar eller geminatar. Etter lange vokalar derimot finn vi det vanlege systemet ved lag, med uaspirerte klusilar som svakt stadium av historiske enkeltlydar, korte preaspirerte klusilar som samanfall av dei to «midtre» gruppene, og lange preaspirerte klusilar i Q3.

Umesamisk tek altså eitt steg vidare jamført med pitesamisk: det er ikkje berre dei to gradane i CC-serien som fell saman etter kort trykksterk vokal, men også sjølve C- og CC-rekkjene. Etter tunge stavingar derimot fungerer umesamisk akkurat som dei fleste andre språka. (Sjå Larsson [2012: s. 122–123] for fleire detaljar om vekslingane i konsonantklynger i umesamisk.) I avsnitt 3 kjem vi tilbake til det spesielle mønsteret i den umesamiske dialekten i Nordre Tärna.

1.3.8. Sørsamisk

Til slutt kjem vi til sørsamisk, som er godt kjend for å mangle synkron stadieveksling.[5] Ser vi på utfalla av dei forskjellige typane klusilar nærmare, finn vi at mønsteret etter korte vokalar er det same som i umesamisk: vi får berre éin type preaspirerte klusilar, uavhengig om dei går tilbake til enkeltlydar eller geminatar, i både svakt og sterkt stadium. Årsaka er også den same, altså at enkeltkonsonantane vart lengde etter kort trykksterk vokal.

Etter lang vokal, derimot, finn vi ein skilnad mellom dei gamle korte og lange klusilane. Førstnemnde får ein *uaspirert* klusil i både sterkt og svakt stadium, medan geminatane har preaspirerte klusilar. Det same skjer når den trykksterke stavinga er tung på grunn av ein kodakonsonant: enkeltklusilar vert reflekterte som uaspirerte og geminatar som aspirerte klusilar eller klynger med stemmelaus sonorant. Sørsamisk har altså ingen motsetnad mellom korte og lange preaspirerte klusilar, noko som umesamisk har (berre etter lange vokalar). Det kan verke paradoksalt at utfallet av dei gamle korte klusilane liknar på deira svake stadium (uaspirerte klusilar) etter lange vokalar, medan det er det sterke stadiet (preaspirerte klusilar) som kjem fram etter korte vokalar; vi kjem tilbake til dette i vår historiske diskusjon nedanfor.

[5] Sørsamisk manglar den såkalla «rotvekslinga», som vi har fokusert på så lenge. Den viser derimot den såkalla suffiksale vekslinga, som er mindre relevant i denne samanhengen.

Kontekst		*p t k		*pp tt kk	
		p̆ t̆ k̆	p̀ t̀ k̀	p̆p t̆t k̆k	p̀p t̀t k̀k
Etter kort vokal	Utfall	hp ht hk			
	IPA	juhkən	juhkə	tihki͡en	tihki͡e
	Ortografi	*johken*	*johke*	*dihkien*	*dihkie*
	Glose	‘elv.GEN’	‘elv’	‘lus.GEN’	‘lus’
Etter lang vokal	Utfall	p t k		hp ht hk	
	IPA	ki͡ətən	ki͡ətə	a:hkan	a:hka
	Ortografi	*gïeten*	*gïete*	*aahkan*	*aahka*
	Glose	‘hand.GEN’	‘hand’	‘bestemor.GEN’	‘bestemor’
Etter sonorant	Utfall	p t k		hp ht hk	
	IPA	juɛ̯lki͡en	juɛ̯lki͡e	kuɻti͡en	kuɻti͡e
	Ortografi	juelkien	juelkie	gurhtien	gurhtie
	Glose	‘fot.GEN’	‘fot’	‘smålom.GEN’	‘smålom’

Tabell 6: Mangel på stadieveksling og klusilar i sørsamisk

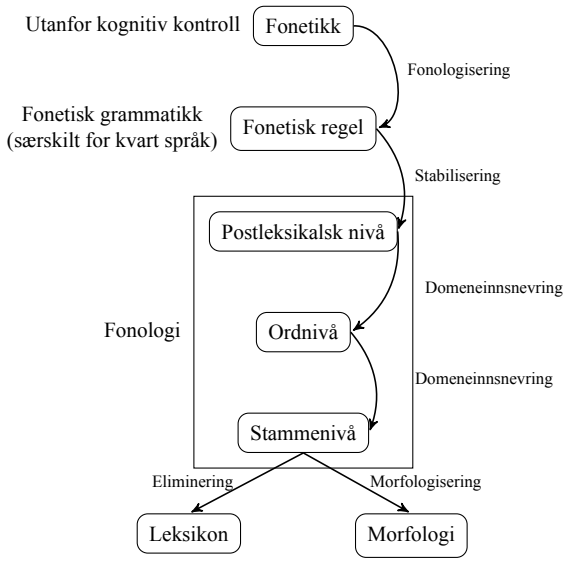
2. Livssyklusmodellen for lydendring

I denne artikkelen bruker eg den såkalla livssyklusmodellen for å forstå korleis lydendringane utviklar seg over tid. Den versjonen som vi tek i bruk her byggjer på arbeida til Kiparsky (1988, 1995) og Bermúdez-Otero (2007, 2015) og Bermúdez-Otero & Trousdale (2012). Figur 1 gir ei grafisk oversikt over strukturen til livssyklusmodellen for lydendringar. To aspekt ved teorien er av spesiell interesse for oss her: vegen som lydendringar tek gjennom ymse område i grammatikken, og korleis nye lydendringar oppstår frå reglar som eksisterer frå før. Vi tek dei i tur og orden.

2.1. Livssyklusen til fonologiske mønster

Ifølgje modellen oppstår fonologiske reglar, som er ein del av grammatikken, frå fonetiske fenomen. I byrjinga er desse fenomena tilfeldige, det vil seie at dei ikkje er noko som talarar bruker, anten på ein medviten måte eller ikkje, som ein del av den fonetiske og fonologiske kompetansen deira. Kjelda til desse fenomena er ofte fysiologiske eller akustiske avgrensingar på kva som er mogleg under taleproduksjon eller -persepsjon; på denne måten er fonologisk endring ofte grunna i fonetiske faktorar (sjå t. d. oversynet hjå Garrett & Johnson 2013).

I denne konteksten kan vi se på preaspirasjon som ein sideeffekt av at uttalen av ein ustemd oral klusil (eller affrikat) inneber to synkroniserte, men separate mekanismar, eller *gestar*: på den eine sida har vi lukking av munnopninga (altså klusilartikulasjon), og på den andre sida må opninga mellom stemmebanda utvidast for å hindre at dei vibrerer når lufta kjem forbi på veg opp. Normalt er dei to gestane synkroniserte, slik at den klangføre artikulasjonen av vokalen stoggar på same tidspunkt som den klanglause orale lukkefasen byrjar. Men viss den gjensidige timinga av dei to gestane ikkje er så tett synkronisert, kan den klangføre artikulasjonen stogge tidlegare eller seinare enn starten av klusilfasen. I det første tilfellet får vi ein periode med stemmeløyse (glottal friksjon) men utan lukking i munnen, altså preaspirasjon (i det andre tilfellet får vi ein kortare periode med klangfør klusil før stemmen stoggar, noko som er vanleg i norske «stemde» klusilar, jf. Ringen & van Dommelen [2013]). Om denne typen preaspirasjon oppstår er avhengig av kor tett timinga er, og denne siste parameteren er noko som er forskjellig i forskjellige språk, sjølv når dei elles har ganske like fonologiske system. Coretta (2020) viser akkurat dette med polsk og italiensk som døme: begge språk har ein kontrast mellom klanglause uaspirerte /p t k/ og klangføre /b d g/, men polsk har ei tettare timing



Figur 1: Livssyklusmodellen for lydendring

mellom oral og laryngal artikulasjon, medan italiensk viser ein tendens til å ha litt tidlegare laryngal opning, som kan føre til «tilfeldig» preaspirasjon (vi kjem tilbake til det italienske dømet seinare).

Over tid kan det skje at ein slik vilkårleg eigenskap vert oppfatta som ein del av korleis ein eller anna fonetisk eller fonologisk kategori vert uttrykt i akkurat det språket. Dette er *fonologisering*, og i første ledd produserer den ein *fonetisk regel*. Dette er eit fonetisk mønster som dirigerer kvantitative aspekt ved realiseringa av ein gitt fonologisk struktur, utan at det faktisk køyrer fonologiske reglar som bytter ut, set inn eller fjernar fonologiske einingar som fonologiske trekk eller autosegmentale assosiasjonar. I denne modellen kan forskjellige språk ha forskjellige fonetiske grammatikkar, som styrer dei finaste detaljane av fonetisk realisering, utan å vise nemneverdige forskjell i den fonologiske strukturen som ligg under fonetikken. Til dømes kan vi førestille oss to språk som har «den same» fonologiske forskjellen mellom kort og lang konsonant. Det finst mange måtar å formalisere ein slik kontrast; vi kan for ordens skyld bruke moraisk teori. Ein fonetisk regel kan seie at ein moraberande klusil kan ha ein preaspirert realisasjon inne i ordet, medan ein ikkje-moraisk klusil ikkje har det. I eit anna språk kan dei same fonologiske representasjonane verte realiserte annleis, til dømes med preaspirasjon hjå både korte og lange klusilar og ein lengdforskjell. På denne måten har dei to språka to forskjellige preaspirasjonsmønster i vidare tyding, men ikkje nokon fonologisk forskjell. Vi skal sjå fleire konkrete døme i diskusjonen av dei samiske språka i avsnitt 3.

Det neste steget er *stabilisering*, der ein fonetisk regel vert til ein fonologisk regel som manipulerer diskrete fonologiske einingar (t.d. segment, fonologiske trekk, stavingsstruktur osv) og ikkje verdiar på tallinja (som formantverdiar, varetid eller tonehøgd). Når det gjeld preaspirasjon så inneber dette steget at preaspirasjon vert til ei sjølvstendig fonologisk eining, oftast ein konsonant som [h]. Etter stabiliseringa kan vi analysere fonologien til språket med reglar som /p^h:/ → [hp]. Denne prosessen er kjent frå både diakrone (‘delinking’ hjå Bye [2001: s. 132]) og synkrone (‘aspiration linearization’ hjå Bals Baal, Odden & Rice [2012: s. 188]) formelle analysar av samisk.

Mykje av arbeidet som er gjort i rammeverket til livssyklusmodellen tek også i bruk ei tilnærming til den grammatiske arkitekturen som går tilbake til *leksikalsk fonologi*. I denne modellen vert den fonologiske komponenten oppdelt i minst to nivå, eller *stratum* (jf. Bermúdez-Otero 2018). Desse stratuma vert definerte på morfologisk grunnlag, og kan ha litt forskjellige fonologiske grammatikkar. Dei fleste modellane skil mellom eit leksikalsk nivå (fonologien «inne i ordet») og eit postleksikalsk nivå, der reglane kan køyre på tvers av ordgrensene. Særs vanleg er ein modell med tre stratum, nemleg to leksikalske (stammenivået og

ordnivået) og eit unikt postleksikalsk nivå. På kvart stratum har den fonologiske grammatikken tilgang berre til materialet som morfologien har «ført til torgs» på det nivået: til dømes kan fonologien på stamme- og ord nivå ikkje operere på tvers av ordgrensene.

Når ein fonetisk regel vert stabilisert, går den inn i den fonologiske grammatikken på det postleksikalske nivået. Deretter går den gjennom *domeneinnsnevring* vidare «ned» til ord- og deretter stammenivået.[6] Deretter kan dei gå heilt ut av grammatikken, men dei kan også verte *morfologiserte*. I dette tilfellet går ein operasjon som byrja livssyklusen som ei lydendring frå å vere ein automatisk regel som køyrer av di den rette fonologiske strukturen er til stades i konteksten, til å fungere som uttrykk for morfologisk struktur.

I det samiske tilfellet er det vel umogleg å halde dei aspekta ved preaspirasjon som er avhengig av den morfologiske strukturen frå det større biletet av stadieveksling, der klusilar inngår som ein relativt liten del saman med dei fleste andre konsonantar. Det er ei kjensgjerning at valet mellom dei tre kvantitetsgradane er tett bunde ved morfologi: til dømes vert stadievekslinga i nordsamisk av både Bye (2005) og Bals Baal, Odden & Rice (2012) analysert som følgje av at eit sett av suffiks vert assosierte med ein flytande mora, som utløyser lengre kvantitetsgradar; jf. også Bye (2007) og Bye, Toivonen & Sagulin (2009) om mønsteret i enaresamisk, som viser eit meir komplekst samspel mellom morfologi og fonologi. Her skal vi sjå bort frå dette aspektet av livssyklusen: etter at den stabiliserte preaspirasjonsregelen har gått inn i systemet, følgjer utviklinga deretter saman med resten av stadieveksling.

2.2. Nye reglar, livssyklusen og dialektgeografi

Eit viktig fenomen som viser interessant samspel med livssyklusen har å gjere med forholdet mellom nye fonologiske reglar og dei som allereie finst i grammatikken. Som vi såg i avsnitt 2.1, ligg ei av kjeldene til nye reglar i fonetisk grunna lydendring. I andre tilfelle kan nye reglar oppstå ved at ein eksisterande regel gjennomgår visse typar av endring. Eitt døme på slike endringar er *generalisering*, når regelen byrjar å gjelde i fleire kontekstar: til dømes kan ein regel som gjaldt berre ein delmengd av alle segment i språket (t. d. berre klusilar) verte utvida til ei større mengd (t. d. alle obstruentar, det vil seie både klusilar og frikativar), eller når den same lydendringa vert utvida til nye kontekstar.

Dette fenomenet har mange namn i litteraturen. Til dømes snakkar ein ofte om «fonetisk analogi» i mange relevante tilfelle (Vennemann 1972). Trass i namnet er det viktig at lydendringar som oppstår gjennom regelgeneralisering eigentleg vanlege lydendringar som går gjennom livssyklusen (altså regulære, *Junggrammatiker*-aktige endringar); jf. diskusjonen hjå Fertig (2013: s. 92–94) om korleis omgrepet «analogi» kan vere misvisande med tanke på regulær lydendring. Bermúdez-Otero (2015), med grunn i Davis (2008), viser korleis den gradvise spreiinga av den høgtyske konsonantforskyvinga gir eit godt døme på korleis ein og same regel vert utvida til stadig fleire kontekstar.

Når ein ny regel oppstår på denne måten, kjem den inn i livssyklusen som alle andre reglar. Samtidig er det ikkje uunngåeleg at den «gamle», meir innsnevra, versjonen av regelen fell ut av grammatikken. Faktisk er det normale utfallet at «nye» og «gamle» variantar av ein fonologisk regel begge finst i grammatikken samtidig, også når den nye har avansert til det neste steget av livssyklusen. Dette fenomenet vert kalla for «spreidde reglar» (*rule scattering*) av Bermúdez-Otero (2015).

På same vis finn vi ofte at (opphavlege) «snevrare» og (nye) «breiare» reglar fortset sin sameksistens i grammatikken. Av di dei breiare reglane er «yngre», er dei ofte eit eller fleire steg attanfor utviklinga til dei eldre reglane. Det kan kome til syne på minst to måtar. For det første kan dei eldre, snevrare reglane vere «djupare» inne i livssyklusen: til dømes kan den eldre versjonen ha gjennomgått stabilisering, medan den nyare enno er på fonologiseringsstadiet (altså fonetisk regel); sjå avsnitt 3.4.3 nedanfor for eit mogleg døme frå umesamisk.

For det andre, ser vi ofte at eldre reglar har ei breiare spreiing i det geografiske rommet, av di dei har hatt tid til å nå fleire varietetar. Denne mekanismen står ofte bak dei klassiske postulata frå historisk

[6] Ikkje alle språk viser klart forskjellen mellom ord- og stammenivå. Spørsmålet har ikkje vore undersøkt i detalj for samisk, så vi legg det til side her.

dialektologi som går tilbake til verk som Schuchardt (1885) og Bartoli (1925), der eldre («arkaiske») trekk vert haldne ved lag i perifere, isolerte område, medan dei meir sentrale, samanhengande sonene viser meir progressive utviklingar. Kristoffersen (2020) viser til eit godt døme på dette frå nordisk språkhistorie. Han tek for seg utviklinga av postvokalisk lenisering («blaute konsonantar»). I den samanhengande «blaute kyststripa» på Sørlandet, i Sør-Sverige og i Danmark er stoda i dag slik at alle norrøne enkeltklusilar vert «oppmjuka» etter vokal (som i Noreg og Sverige er visst lang på grunn av kvantitetsforskyvinga), altså *pibe*, *bide*, *bage*. Kristoffersen (2020) peikar ut eit mønster der lenisering berre har skjedd etter opphavleg, ikkje lengd kort vokal, med former som *kjødd* og *sidde* (norrønt *kjǫt*, *sitja*) men *bite* (< *bíta*), *bake* (< *baka*), *sjip* (< *skip*). Eg er samd med han at slike tilhøva er mest kompatible med eit scenario der lenisering først skjer etter korte vokalar; den meir ekspansive regelen oppstår seinare, og når dimed ikkje dei fjernaste strøka i leniseringsområdet. Då er det spesielt slåande at dette meir arkaiske mønsteret med lenisering berre etter kort vokal finst i to separate soner, som i begge tilfelle ligg på kanten mellom eit samanhengande storområde med lenisering og den ikkje-leniserande sonen: Sandøya-området aust for Arendal mot fylkesgrensa med Telemark, og Iddedalen i Østfold, på grensa mot den «blaute» stripa langs Bohuslänskysten.[7]

Desse mekanismane vil vere viktig for vår diskusjon, av di eg hevdar at dei gir oss særs gode heuristiske kriterium for å spore utviklinga av preaspirasjon gjennom dei samiske språka.

3. Utviklinga til preaspirasjon i samiske språk

I dette avsnittet drøftar eg korleis vi kan forstå opphavet og utviklinga til preaspirasjon i samiske språk i lys av livssyklusmodellen som skissert i avsnitt 2. Eg argumenterer for at denne modellen er særs eigna til å gi oss ny innsikt i variasjonen som vi finn i materialet.

3.1. Opphavet til stadieveksling

I og med at det uralske urspråket mangla (iallfall distinktive) preaspirerte klusilar, må vi rekne med at preaspirasjon er ein innovasjon som oppstod på eit tidspunkt mellom ururalsk og dagens språk. Av di preaspirasjon heng så tett saman med stadieveksling, er det naturleg å ty til akkurat stadievekslinga for å forstå kvar preaspirasjonen kjem frå.

Kva er altså opphavet til stadieveksling? Det verkar ganske klart at den er, i grunnen, eit kvantitativt fenomen. I dei fleste språka innom både den austersjøfinske og den samiske greina ser vi at svakt stadium er på ein eller annan måte «kortare» enn sterk stadium. Dette kjem klarast fram i geminatrekkja, til dømes i det utbreidde samiske systemet med lang preaspirasjon i *ČC vs. kort preaspirasjon i *ČC. I enkteltlydrekkja er utfallet mindre klart, av di dei fleste austersjøfinske og nokre samiske språk (spesielt i aust) viser såkalla *kvalitativ veksling*, der *Č-rekkja ikkje er representert ved klusil (av typen estisk *jõgi* ‘elv.NOM’ ∼ *jõe* ‘elv.GEN’, SaaN *goahti* ‘gamme.NOM’ ∼ *goađi* ‘gamme.GEN’). I tillegg får vi i austersjøfinsk eit samanfall mellom den svake graden av enkeltklusilen **t* (med utfall som [d], [r], [l], [j] i svakt stadium) og ururalsk **δ* (som godt kunne vere [ð]), t. d. F *kato* ‘hus’, *kadon* ‘hus.GEN’ med **t* (ungarsk *ház*), F *pato* ‘dike’, *padon* ‘dike.GEN’ med **δ* (ungarsk *fál* ‘vegg’).

Vi kan sjå for oss to moglege tolkingar av prosessen som endar med eit «svakare» og eit «sterkare» utfall: anten vi har å gjere med ein *lenisering*, altså at lydane som vert utsette for prosessen vert «veikare», eller så har vi ei *styrking*. Tradisjonelt vert stadieveksling sett på som lenisering (t. d. Wiklund 1896, Posti 1953, Ravila 1960, Korhonen 1981, 1988). Blant anna gir dette gode resultat i austersjøfinsk, der kvalitativ veksling, altså ikkje-klusilar i svakt stadium, er vanleg. Frikativar eller sonorantar er naturlege produkt av lenisering, men det er ikkje veldig ofte at vi ser styrking av frikativar til klusilar (sjå om dette Bybee & Easterday 2019): altså viss kvalitativ veksling, som er godt spreidd i austersjøfinsk og også til stades i samisk, er gammal, då gir lenisering eit godt scenario for heile vekslinga.

[7] I Iosad (under utarb.) legg eg fram ein meir inngåande diskusjon om leniseringsmønstera i nordiske språk. Sjå også Ramsammy (2018) for eit liknande døme frå den såkalla *gorgia* i italienske dialektar.

Sett frå samisk hald er denne idéen derimot problematisk (sjå allereie Bergsland 1945). Til forskjell frå dei austlege varietetane har språk som lulesamisk eller sørsamisk klusilane overalt, og det finst inkje prov for tidlegare frikativar i stavingsveksling. Samisk manglar også samanfall mellom **t* og **δ*: i språk som nordsamisk fell dei saman som [ð] i svakt stadium (*goađi* 'hus.GEN.SG', *buođu* 'demning.GEN.SG') men er fortsett distinkte i sterkt stadium (*goahti* 'hus.NOM.SG', *buođđu* 'demning.NOM.SG'). I sørsamisk er desse to konsonantane heilt distinkte (*gåetie* men *buore(ve)*), noko som talar for at kvalitativ veksling kan berre vere sekundær i samisk.

Viss vekslinga i utgangspunktet er kvantitativ, vert det vanskelegare å skilje mellom lenisering og styrking som grunnlag for utviklinga. Blant anna er det ganske vanskeleg å sjå sambandet mellom endringa (lenisering eller styrking) og konteksten for vekslinga, nemleg om den påfølgjande stavinga er open eller lukka. Her går eg ut ifrå at forfattarar som Gordon (1997), Sammallahti (1998, 2012) og Bye (2001) har rett når dei ser på vekslinga som, i grunnen, ei styrking (jf. også Bańczerowski 1969).

Eitt argument for dette, som bl. a. Gordon (1997) tek for seg, er at både dei austersjøfinske og dei samiske språka viser uomtvista døme på styrking av konsonantar framfor lange vokalar i utviklinga av vekslingssystemet. Denne typen lenging er *ikkje* den same som dei vi har sett så lenge på. Den ligg til grunn for overlange konsonantar (Q3) framfor lange vokalar, som oppstår blant anna som følgje av konsonanttap, jf. estisk *kätte* 'hand.ILL' < **kätēn* < **käte-hen* (F *käteen*). Dette fenomenet finn vi i estisk, ingrisk og samisk, og det er denne lenginga som gir overlange frikativar eller sonorantar, som i nordsamisk [sul:lo] *sul'lo* 'øy.GEN' (**suolōjin*; jf. NOM. *suolu* < **suolōj*).[8]

Sjølv om dei to typane lenging ikkje er identiske, fremjar Gordon (1997) hypotesen om at den «vanlege» vekslinga i grunnen tilhøyrer den same typen: viss vokalar i den påfølgjande stavinga er lengre når denne stavinga er open, vert også konsonantane lengde, og dimed inntrer den kvantitative vekslinga som vi kjenner den. Ei viktig følgje frå denne hypotesen er at det samiske systemet, der vekslinga famnar både klusilar, frikativar, sonorantar, og konsonantklynger, er det opphavlege. Til inntekt for dette tek Gordon (1997) det faktum at tendensen til lenginga av alle konsonantar framfor lange vokalar finst også i finsk, som elles berre har veksling av klusilar (jf. Suomi, Toivanen & Ylitalo 2008: s. 90). Viss det meir ekspansive systemet sanneleg er det eldre, må mønsteret ha vore *innskrenka* i dei fleste austersjøfinske språk til å gjelde berre klusilar, med berre dei fonetiske tendensane som spor av den tidlegare stoda.

Under livssyklusmodellen stemmer dette scenarioet mest sannsynleg ikkje. Dei finske dataa er verdifulle, av di dei påviser at tendensen til lenging av konsonantar framfor lengre vokalar er ganske allment i den finno-ugriske språkgreina, men ein fonetisk tendens er normalt ei kjelde til ein fonologisk regel, ikkje eit spor. Eg slår dimed følgje med forfattarar som Bergsland (1945) og Ravila (1960), som ser «fonetisk» veksling som eit tidleg fenomen, kan hende sams for austersjøfinsk og samisk, og «fonologisk» veksling som ei seinare utvikling.

Opphavet til stadieveksling er dimed ein fonetisk tendens til lenging av konsonantar framfor vokalar i opne stavingar. Vi skal ikkje gå nærmare inn på kvar denne tendensen kjem frå. Tradisjonelle forklaringar går ut på skilnadene i sterkare eller svakare «aksent», men vi kan for det meste avskrive dei som «papirfonetikk». I mykje av den generativ-fonologiske litteraturen vert problemstillinga drøfta frå ein innfallsvinkel som tek for seg metrisk struktur. Ei lang andrestaving gir ikkje nokon god fotstruktur når fotstrukturen i uralske språk er til vanleg trokaisk, slik at den første stavinga i ein metrisk fot bør vere minst like tung (ideelt tyngre) enn den andre; lengd i andrestavinga utløyser dimed ei lenging i førstestavinga. Denne «metriske optimiseringa» er veldig utbreidd i den seinare historia til både austersjøfinske og samiske språk (jf. Gordon 1997, Bye 2005, Kiparsky 2008, 2018). Det viktigaste her er likevel at *alle* konsonantar må ha vore omfamna av regelen, både klusilar/affrikatar, frikativar, og sonorantar. Viss årsaka var metrisk optimisering, så er det absolutt ingen grunn til å ekskludere nokon av konsonanttypane frå trongen til lenginga som står bak fenomenet, og sidan alle språka viser lengingstendensen i ei eller anna form, er det nærliggjande å rekonstruere denne stoda heilt tilbake i tida.

Under livssyklusmodellen må vi rekne med at denne tendensen til styrking framfor ei lengre staving

[8] Fleire dialektar, som t. d. Eanodat, viser også andre lengingsfenomen (jf. Sammallahti [1998: s. 49] om «lenging»); dette mønsteret er også interessant, men av litt mindre relevans i denne konteksten (takk til den anonyme fagfellen for avklarande diskusjon).

byrja som ein fonetisk regel i dei austersjøfinske og samiske urspråka, av di denne tendensen på ingen måte er universell. Den fonetiske regelen styrer altså varetida til konsonantar, utan å lage ein skilnad i fonologisk representasjon. Dimed har prosessen gått gjennom det første steget i livssyklusen, nemleg *fonologisering*. Denne utviklinga må ha vore sams for dei to urspråka.[9]

I begge to har det vidare ført til *stabilisering*, i første omgang når det gjaldt klusilane og affrikatane. Dette skapa ein *fonologisk regel*, der både korte og lengre klusilar fekk kortare allofonar framfor lukka stavingar, og lengre allofonar framfor opne stavingar. I den grammatiske arkitekturen som vi tek i bruk her tyder dette at svake og sterke allofonar av dei to klusilrekkjene var skilde frå kvarandre på ein måte som innebar ein skilnad i den fonologiske representasjonen deira. Kva denne skilnaden var er ikkje så lett å seie: systemet bør vere i stand til å skilje mellom korte og lange geminatar, noko som er veldig kontroversielt (jf. Bye 1997, 2001, Odden 1997, Bals Baal, Odden & Rice 2012, Prillop 2013).

Det er iallfall rimeleg å rekne med atRavila (1960), Sammallahti (1998, 2012) og Bye (2001) har rett når dei ser på prosessen i det minste i samisk som ei styrking framfor opne stavingar. Det er fullt mogleg at reglane er forskjellige i samisk (styrking) og austersjøfinsk (lenisering). Blant anna gir denne tilnærminga ei god avklaring av åtferda til konsonantklynger, som t. d. Ravila (1960) peikar på. I finske klynger av typen *lk* vekslar klusilen på akkurat same måten som etter vokal (*jalka* ‘fot.NOM.SG’ ∼ *jalan* ‘fot.GEN.SG’). I samisk, derimot, er det vanleg at det sterke stadiet inneber ei lenging av den første konsonanten (SaaP [ɲa:r:ka] *njárrga* ‘nes.NOM.SG’ ∼ [ɲa:rka] *njárga* ‘nes.GEN.SG’), medan klusilen kan faktisk verte sterkare i det svake stadiet (som i SaaN [pe͡alk:i] *bealggi* ‘tommel.GEN’, svakt stadium av [pe͡alĕgi] *bealgi* ‘tommel’). Leniseringa av klusilen gir dimed inga avklaring for det samiske mønsteret.

I samisk, men ikkje i austersjøfinsk, finst ein versjon av denne fonologiske stadievekslingsregelen også i sonorantar.[10] I livssyklusmodellen er dette eit praktdøme på korleis ein regel vert generalisert: konteksten for regelen vert enklare, og den omfamnar fleire konsonantar enn før. Alternativet, forfekta av Gordon (1997), er at vekslinga hadde vore fonologisert for både klusilar og sonorantar, men vart seinare missa i sistnemnde. Dette er strengt teke mogleg: fonologisering inneber ikkje at den tilsvarande fonetiske regelen døyr ut (*rule scattering*), og skulle den fonologiske vekslinga i sonorantar ha gått tapt i austersjøfinsk, ville den fonetiske regelen kome til syne igjen. Dette scenarioet er likevel mykje meir innfløkt enn det meir økonomiske alternativet med berre éin innovasjon av ein ganske vanleg type.

Der stadieveksling ikkje vart stabilisert vidare enn til klusilar, må vi likevel rekne med at den fonologiserte regelen som lengde alle konsonantar i visse kontekstar likevel varte ved i språket. Vi ser dette både ved at den fonetiske tendensen har vore halden ved lag i fleire språk, og ved at seinare døme av metrisk optimisering med røter i same fenomenet dukka opp fleire gongar seinare i språkhistoria, som nemnt ovanfor.

Ein anonym fagfelle spør om ikkje stadieveksling kunne sjåast på som ikkje som ein reaksjon mot «for mykje» vekt i andrestavinga, der den første stavinga vert tyngre for å oppretthalde forholdet mellom dei to, men heller eit kompenseringsfenomen utløyst av at andrestavinga vert redusert når ho er open, men ikkje når ho er lukka: jf. tersamiske former som *jogg* ‘elv.NOM’ (sterkt stadium, redusert kort vokal i open staving) men *joga* ‘elv.GEN’ (svakt stadium, bevart vokal i opphavleg lukka staving). Slike reduksjonsfenomen finst i fleire språk, blant anna i mykje av austsamisk, estisk, livisk, og ingrisk. Under dette scenarioet inngår også alle konsonantar i vekslingsmønsteret, liksom i mitt framlegg, men veksling vert seinare innsnevra til klusilar i austersjøfinsk, kan hende av di dei andre konsonanttypane ikkje er fonetisk godt eigna til denne kompenseringsfunksjonen. Dette er eit interessant framlegg, men empirisk er det vel eit problem at vi finn stadieveksling i beste velgåande også i språk som finsk eller nordsamisk, som ikkje har mykje vokalreduksjon. Eg vil også understreke at det er sjølve innsnevringsomgrepet som er problematisk sett frå eit livssykklusperspektiv: at ein fonologisk regel «bryt ned» til eit fonetisk mønster er rett og slett ikkje noko som modellen tillèt. På denne måten er dette ei klar føreseiing av modellen at utviklinga ikkje ha gått frå fonologisk regel til fonetisk tendens, uavhengig av den opphavlege motivasjonen.

[9] Her kan vi ikkje gå inn nærmare på den djupare historia. Dei siste resultata i uralistikken avviser eit sams finsk-samisk urspråk (t. d. Aikio 2015, Zhivlov 2015). På den andre sida har Helimski (1996) gjenoppliva den gamle hypotesen om stadieveksling som fellesuralisk fenomen, og det er godt mogleg at ei ny tilnærming basert på livssyklusmodellen kan vere nyttig her.

[10] I austersjøfinske språk der stadievekslinga omfamnar sonorantar, som estisk og ingrisk, er dette generelt ei følgje av dei seinare lengingsprosessane, ikkje av den opphavlege «rotvekslinga».

3.2. Kvar kjem preaspirasjonen frå?

Enn så lenge har vi tala om opphavet til stadieveksling, ikkje til sjølve preaspirasjonen. Kvifor får vi preaspirasjon som eit så gjennomgåande trekk i vekslingssystemet?

Her er det særs viktig at scenarioet vår for stadievekslinga inneber at den i grunnen er eit kvantitativt fenomen. Som vi såg ovanfor manglar det prov i samisk for kvalitativ veksling tidleg i språkhistoria. Frå før hadde språket ein kvantitativ kontrast mellom enkeltlydar, og fonologiseringa av stadievekslinga førte til at det kunne finnest opp til fire distinkte fonologiske representasjonar med lengdforskjell.[11]

Det er godt kjent at fonologiske kontrastar ofte vert «opptrappa» (engelsk *enhancement*; jf. Stevens & Keyser [1989, 2010] og Hall [2011]), ved at fonetiske eller fonologiske trekk som ikkje er elles distinktive i språket vert «rekrutterte» for å signalisere kontrastane betre. Kvifor var det preaspirasjon som vart til denne typen «hjelpereiskap» i kvantitetssystemet?

Samisk skil seg ut frå andre språk i Nordeuropa som har preaspirasjon ved at språka mangla anten aspirerte klusilar eller fonemet [h] på tida då preaspirasjonen oppstod. Både germanske og keltiske språk har som regel begge to (sjå t. d. Salmons [2020] om germansk og Eska [2018] om keltisk), medan ursamisk mangla laryngale kontrastar i klusilar i det heile teke, og hadde heller ikkje noko *h*-segment. Den nærmaste parallellen til denne stoda finn vi i italiensk (t. d. Gobl & Ní Chasaide 1999, Stevens & Hajek 2007, Stevens 2011, Stevens & Reubold 2014, Coretta 2020). Dette er eit anna språk som manglar aspirasjon og fonemisk [h], men som kan oppvise preaspirasjon i stemmelause klusilar, spesielt i geminatar, akkurat som i samisk.

Vi har allereie nemnt italiensk som eit anna språk der den fonetiske førelauparen av preaspirasjon er til stades, i form av eit laust timingsforhold mellom den laryngale opninga og klusilfasen. Dimed har vi eit akustisk signal som av og til opptrer med stemmelause klusilar,[12] men vert ikkje brukt til noko anna i språket. Det er dimed tilgjengeleg som opptrappingsreiskap for ein kontrast som klusilane tek del i. Den detaljerte studien av Stevens & Reubold (2014) viser at preaspirasjon i italiensk aukar den totale varetida på VC-sambandet, utan at det verkar inn på persepsjonen av kvantiteten til sjølve klusilen som kort eller lang.

Eg legg dimed fram at preaspirasjon i samiske språk oppstod først som eit tilfeldig, ikkje-kontrollert fenomen i samband med artikulasjonen av stemmelause klusilar. Deretter vart den rekruttert først som ein fonetisk, kontrollert regel for å signalisere dei kvantitative tilhøva i ordet, av di den eignar seg betre til dette enn rein varetid. Denne fonologiseringa er ganske lik den som Pétur Helgason (2002) la fram for nordiske språk: forskjellen ligg i at nordiske stemmelause klusilar var aspirerte frå før, og der var preaspirasjon dimed ein type koartikulasjon mellom kontrastiv aspirasjon og vokalen som kom framfor konsonanten, medan i samisk oppstod den i samband med kvantitative kontrastar.

Dimed reknar vi med at preaspirasjon gjennomgjekk det første steget i livssyklusen, nemleg fonologisering. I det neste avsnittet ser vi på nokre vidareutviklingar.

3.3. Når oppstod preaspirasjonen?

Som vi såg i avsnitt 1.3 har dei aller fleste samiske språka to rekkjer preaspirerte klusilar: klusilar med lang preaspirasjon (som vi finn spesielt i sterkt stadium av historiske geminerte klusilar) og dei med kort preaspirasjon, som ofte, men ikkje alltid, svarar til både geminatar i svakt stadium og enkeltklusilar i sterkt stadium. For å finne ut kronologien til korleis preaspirasjonen utvikla seg, bør vi no granske tilhøva mellom dei samanfalla som vi finn i vekslingssystemet og preaspirasjon. Med andre ord: viss vi finn at to eller fleire grader av vekslinga fell saman i eit gitt språk, oppstod preaspirasjon før eller etter dette samanfallet? Av spesiell interesse er tilfelle der vi kan vise at preaspirasjon utvikla seg *etter* eit samanfall, av di dette kan vere prov på ei utvikling i preaspirasjonsmønster som ikkje går heile vegen tilbake til ursamisk.

[11] Vi tek ikkje opp spørsmålet om alle fire stadium kunne vere til stades i språket samtidig; sjå drøftinga i Bye (2001: s. 130–132), som er godt kompatibel med framlegga i denne artikkelen.

[12] Eller stemmelause segment meir generelt. Preaspirasjon finst i nokon mon faktisk også framfor stemmelause frikativar, jf. om engelsk Hejná (2015). Den er også notert i detaljerte transkripsjonar av samisk material, sjå Bańczerowski (1969: s. 139–141) for ei stutt drøfting. Den vert likevel sjeldan stabilisert på same vis som framfor klusilar.

	*p t k		*pp tt kk	
	p̆ t̆ k̆	p̀ t̀ k̀	p̆p t̆t k̆k	p̀p t̀t k̀k
Urform	*kātōm	*kātōtēk	*kāttōn	*kāttō
Nordre Tärna	ˈkaːtu͡op	ˈkaːhtu͡ot	ˈkaːhˑtu͡on	ˈkahːtu͡o
Lulesamisk	*gádov*	*gáhtot*	*gáhto*	*gáhtto*
Glose	‘vere.borte:PRS.1SG’	‘vere.borte:INF’	‘katt.GEN.SG’	‘katt.NOM.SG’

Tabell 7: Fire vekslingsgrader i Nordre Tärna

Det «vanlege» mønsteret der C̀- og C̆C-rekkjene fell saman som preaspirerte klusilar er diverre ikkje så informativt i seg sjølv: vi kan godt sjå for oss at samanfallet skjer først, og konsonantane i den samanslegne rekkja utviklar preaspirasjon seinare, men like godt kunne det ha vore tilfellet at éi av rekkjene fell saman med ei rekkje preaspirerte klusilar som hadde oppstått tidlegare.

Det som er viktig er derimot at sjølve samanfallet av **p̀ t̀ k̀* og **p̆p t̆t k̆k* ikkje er av ursamisk dato. Det finst tre prov for det. For det første, kjem dette ur av det pitesamiske mønsteret der C̆C- or C̀-rekkjene vert handsama på forskjellig vis etter kort vokal: C̆C, men ikkje C̀, fell saman med sterkt stadium av geminatar (C̀C), noko som fører til at stadieveksling av geminatar fell bort i denne konteksten. Om vi tilskriv dette ei lydendring og ikkje analogisk nivellering av vekslinga (som nok er rett, jf. drøftinga nedanfor) så må dette samanfallet ha funne stad før samanfallet av **p̀ t̀ k̀* og **p̆p t̆t k̆k* (som vi finn etter ikkje-kort vokal), elles hadde også dei gamle sterke enkeltklusilane teke del i det.

Det andre, litt utfordrande, argumentet kjem frå dei kolasamiske språka. Som vi drøfta i avsnitt 1.3.1 viser iallfall nokre kjelder eit mønster der C̀-klusilane vert reflektert som stemde geminatar, medan konsonantane i C̆C-rekkja vert realiserte som ikkje-preaspirerte ustemde enkeltklusilar. Situasjonen der er ikkje så klart, men viss dette stemmer så har vi eit anna døme utan dette samanfallet (og utan preaspirasjon i C̀-rekkja i det heile teke, noko som elles finst overalt i samisk).

Ei særs interessant stode fanst i den no utdøydde umesamiske dialekten i Nordre Tärna, som vert diskutert av Bergsland (1973) på grunn av ei upublisert skildring av Moosberg (1920) og drøfta bl. a. av Sammallahti (1998, 2012) og Bye (2001).[13] Ifølgje Bergsland (1973) skilde denne dialekten fortsett mellom alle fire konsonantgrader etter ikkje-korte vokalar, med uaspirerte klusilar i svakt stadium av enkeltklusilar og så tre grader av aukande lengd på preaspirasjonen. Tabell 7 (etter Bergsland 1973, Sammallahti 2012) viser dette. Den same mangelen på samanfall og eit firedelt mønster fanst med nasale konsonantar òg i fleire umesamiske dialektar (Sammallahti 1998: s. 194, Larsson 2012: s. 120–121).

Alt dette viser at samanfallet av dei to «midtre» gradane i vekslingssystemet ikkje går tilbake til ursamisk. Det einaste preaspirasjonsmønsteret som *alle* samiske språk (utanom enaresamisk; sjå nedanfor) oppviser er (lang) preaspirasjon i sterkt stadium av opphavlege geminerte klusilar. Dei aller fleste har også preaspirasjon i svakt stadium av geminatrekkja. Det er berre på Kola, i kildin- og tersamisk (og kan hende i akkalasamisk også), at C̆C-rekkja manglar preaspirasjon etter vokalar. Dette stemmer både om C̆C-konsonantane fell saman med C̀-rekkja som stemde geminatar, eller om dei vert handsama som uaspirerte ustemde klusilar. Men som vi allereie har vore inne på, er denne mangelen sannsynlegvis ei sekundær utvikling (Sammallahti 1998): etter sonorantar er gamle geminatar preaspirerte i både sterkt (SaaK *toorrhk* ‘pelskåpe’, *lee´mmhk* ‘skulderreim’) og svakt stadium (*torhk* ‘pelskåpe.GEN.SG’, *le´mhk* ‘skulderreim.GEN.SG’). Sjølv om det kan verke uøkonomisk å gå ut ifrå eit seinare tap av preaspirasjon etter vokalar, viser dette mønsteret klart at preaspirasjon må ha vore til stades også hjå svake geminatar.

Sammallahti (1998: s. 193, 195) sin interpretasjon av denne stoda er at preaspirasjon av geminerte klusilar var ei ursamisk lydendring. At dei sterke enkeltklusilane fell saman med korte preaspirerte klusilar er ei seinare utvikling, som han listar opp under «urnordvestsamisk» (altså urspråket til pite-, lule- og nordsamisk). Vi finn samanfallet også i (dagens) umesamisk (etter lange vokalar) og i fleire austlege dialektane: definitivt

[13] Larsson (2012), som også byggjer på Moosberg sitt arbeid, drøftar ikkje desse tilhøva i detalj.

Språk	*pp > ʰp	p̀ > ʰp	Rotveksling	Q2-samanfall	p̆p > p̀p / V̆_	p > pp / V̆_
Kolasamisk	⊘		✔	⊘		
Skoltesamisk	✔	✔	✔	✔		
Enaresamisk	⊘	⊘	✔	✔		
Nordsamisk	✔	✔	✔	✔		
Lulesamisk	✔	✔	✔	✔		
Pitesamisk	✔	✔	✔	✔	✔	
Umesamisk	✔	✔	✔	✔	✔	✔
Tärna-umesamisk	✔	✔	✔		✔	✔
Sørsamisk	✔	⊗			⊗	✔

Tabell 8: Utviklinga av preaspirasjon og stadieveksling i samiske språk

i enare-, skolte- og akkalasamisk. Utviklinga i kildin- og tersamisk er, som vi har sett, ikkje heilt avklarte, og ifølgje Sammallahti (1998) sannsynleg sjølvstendig, sjølv om dei iblant kan likne på dei vestlege. Eit liknande scenario, med (minst) to «bølgjer» av preaspirasjon — først i geminatar og deretter i enkeltklusilar — er lagt fram av Bye (2001: s. 132). Eg seier meg samd med dei. Vidare er det verdt å påpeike at ei slik utvikling mogleg gir oss nok eit anna døme av regelgeneralising: systemet går frå ein regel om preaspirasjon i ein delmengd av klusilane (geminatane i dette tilfellet) til ein regel som famnar fleire klusilsegment.

Denne forståinga av samspelet mellom stadievekslinga og preaspirasjon kastar også lys over opphavet til mønster med manglande synkron stadieveksling. Desse finn vi i pitesamisk (hjå geminatar etter kort trykksterk vokal), umesamisk (alle konsonantar etter kort trykksterk vokal) og sørsamisk (i det heile teke). Den tradisjonelle forklaringa av denne mangelen (t. d. Wiklund 1896, Collinder 1929) er analogisk nivellering, moglegvis under germansk innverknad. Problemet med denne teorien er at sørsamisk konsekvent viser det *sterke* stadiet som utfall av nivelleringa blant gamle korte klusilar etter korte vokalar (*johke* ‘elv’), men det *svake* etter lange vokalar (*gïete* ‘hand’). I sin klassiske artikkel viser Bergsland (1945) at ei betre forklaring er at stadieveksling aldri vart fonologisert i sørsamisk i same form som i dei andre språka: mønsteret er forklart om gamle geminatar vart lengde etter kort trykksterk vokal (altså i **jokę* ‘elv’ > **jokkę* men ikkje i **kētę* ‘hand’), og så fekk alle geminatar, både opphavlege og sekundære, preaspirasjon. Samanfallet mellom **p̀ t̀ k̀* og **p̆p t̆t k̆k* fann aldri stad, korkje etter kort eller lang vokal. Dette forklarar både mangelen på stadieveksling og mangelen på ein kontrast mellom lange og korte preaspirerte klusilar i sørsamisk.

Viss dette er den rette løysinga (og det trur eg at det er) så er det attraktivt å forklare også andre tilfelle av manglande synkron stadieveksling med regulær lydendring heller enn seinare tap (analogisk nivellering); jf. om dette også Ravila (1960). Som vi så ovanfor har vi tre forskjellige mønster på slik mangel: hjå geminatar etter korte vokalar (pitesamisk), hjå alle konsonantar etter korte vokalar (umesamisk) og hjå alle konsonantar etter alle vokalar (sørsamisk).

Vi kan arrangere også desse reglane i ein annan rekkje av regelgeneralisering. Tabell 8 viser kva for nokre av dei lydendringane eg rekonstruerer som er til stades i kvart enkelt språk. Dei første spaltene viser om noka form for preaspirasjon finst i språket som utfall av geminatar (som vi har sett skil ingen av språka mellom sterke og svake geminatar med tanke på preaspirasjon, med eit delvist unnatak i dei kolasamiske varietetane) og av sterke enkeltklusilar. «Rotveksling» viser om språket skil mellom sterkt og svakt stadium i minst éin kontekst. Spalta «Q2-samanfall» viser om C̀-rekkja er identisk med C̆C-rekkja: som vi har sett, er også dette litt problematisk på Kola, men sjølv om samanfallet finst der, er resultatet av det *ikkje* preaspirert. Dei to siste spaltene viser konsonantlengingsreglane som vi nettopp har peika ut som del av ei generaliseringsrekkje. Språka er grovt arrangert frå nordaust til sørvest. Symbolet ⊘ tyder at trekket må ha vore til stades i språket tidlegare, men har i dag ei anna form. Sørsamisk er vist separat av di fleire av endringane (nemleg dei som inneber ein skilnad mellom sterkt og svakt stadium) ikkje kan ha skjedd der sidan dette språket aldri fekk rotveksling til å byrje med.

Preaspirasjon av opphavlege geminatar er sams for alle samiske språk. Dette kan vere, som sagt, opp

til ei ursamisk lydendring, men fenomenet kan også ha spreidd seg som ein tidleg innovasjon. Det er ikkje lett å seie kor gammalt det er. På den eine sida kan det ha oppstått allereie før dei samiske språka spreidde seg til det nordlege Fennoskandia: Aikio (2012) peikar ut *Kuhkaa*, namnet på ei avlang øy i Ladogasjøen, som han jamfører med PSaa **kukkē* 'lang' (SaaN *guhkki*). På den andre sida var preaspirasjon av geminerte klusilar fortsett aktiv som ein regel i språket då sørsamisk gjennomgjekk lenginga av opphavlege enkeltlydar etter korte trykksterke vokalar. Viss dette fenomenet verkeleg har noko å gjere med dei einsarta utviklingane i nordisk og spesielt i trøndsk, kan vi dagsetje denne perioden til seinmellomalder.[14]

Tabell 8 viser vidare at preaspirasjon av sterke enkeltklusilar var den neste innovasjonen. Han ser ut til å ha hatt eit fokus i eit nordvestleg område. I aust nådde han ikkje Kola,[15] og i sør finn vi ingen spor av fenomenet i sørsamisk (som i første omgang ikkje skil mellom sterkt og svakt stadium). At preaspirerte utfall av sterke enkeltklusilar ikkje skyldest det generelle samanfallet av C̀- og C̆C-rekkjene kan vi sjå frå mønsteret i den umesamiske dialekten i Nordre Tärna, som hadde preaspirasjon i sterk grad av enkeltklusilar men ikkje noko Q2-samanfall. Vi kan dimed slutte at preaspirasjon i C̀-serien var ein tidlegare innovasjon enn samanfallet, og at den geografiske distribusjonen speglar den temporale dimensjonen.

Denne spreiinga i sørleg retning møtte innovasjonar som innebar nøytralisering av stadiekontrastane gjennom konsonantlenging. Den tidlegare innovasjonen, som nesten når den nordlege grensa av det pitesamiske området, er samanfallet av sterkt og svakt stadium av geminerte klusilar. At dette ikkje skyldest ei generell lenging kan vi sjå ved at forskjellen mellom sterke og svake geminatar står ved lag etter ikkje-kort vokal. På det neste steget vert den generalisert til å gjelde også korte konsonantar etter kort vokal; denne regelen er litt yngre, og har berre nådd fram til umesamisk. Det er sannsynleg at innovasjonen var utløyst av mangelen på kvantitative kontrastar i sør, men distribusjonen in tabell 8 viser at prosessen gjekk nordover på ein gradvis måte.

I dette avsnittet har eg prøvd å vise korleis vi kan bruke livssyklusmodellen for å avsløre den interne logikken og kronologien til utviklinga av preaspirasjon og den innbyrdes påverknaden mellom den og stadievekslingssystemet. To generelle moment er viktige her:

- Viss vi ser vi på utviklingane som ein serie av regelgeneraliseringar heller enn ei stor endring eller ein haug isolerte utviklingar, så kan vi spore både den relative kronologien til korleis systemet oppstod og den geografiske systematikken i mønstera.

- Det sentrale konseptet i livssyklusmodellen er at nye fonologiske reglar oppstår frå fonetisk variasjon i urspråket. Dei eksisterande reglane kan også spele ei rolle, idet dei kan tene som føredøme til generalisering. Dette forklarar kvifor vi finn ei rekkje liknande, men samtidig forskjellige lydendringar som går igjen i historia til det same språket, eller i fleire nærståande språk (i dette andre tilfellet snakkar me ofte om *drift*, etter Sapir [1921]): dette skjer av di alle disse endringane kjem frå den same fonetiske variasjonen, og vert drivne av livssyklusen på liknande måtar.

I dette avsnittet la eg fram korleis utviklinga av preaspirasjon i samiske språk er ein god illustrasjon for begge desse poenga. I det neste skal vi sjå på korleis preaspirasjon utvikla seg langs livssyklusen.

3.4. Preaspirasjon i samisk: fonetikk eller fonologi?

Ifølgje livssyklusmodellen bør preaspirasjon oppstå som ein fonetisk regel, men deretter kan den utvikle seg til eit fonologisk fenomen. Her skal eg vise at det er akkurat dette som skjer med samisk preaspirasjon. For å gi drøftinga eit teoretisk grunnlag byrjar vi med ein kort diskusjon av kriteria som vi kan bruke for å skilje mellom dei to.

[14] Liknande utviklingar, der lenging av gamle enkeltklusilar «nærer» preaspirasjon av geminatar, er godt kjende i nordisk frå herjedalsmåla (Reitan 1930).

[15] Sammallahti (1998: s. 55) legg også fram at enaresamisk og skoltesamisk fekk systemet deira under innverknad frå nordvestsamiske språk.

3.4.1. Modularitet og livssyklusen

I avsnitt 2.1 lå eg fram ein modell for både synkron og diakron fonologi som gjer skilnad mellom *fonetiske* og *fonologiske* reglar. For å skilje mellom dei skal vi bruke *modularitet* som hovudkriterium. Det vil seie at fonetikk og fonologi er to distinkte grammatiske modular, som manipulerer forskjellige typar einingar og er opne for forskjellige typar påverkande faktorar. Ovanfor så vi at fonetiske reglar styrer verdiar som er uttrykt i reelle tal, medan fonologiske reglar manipulerer diskrete einingar som segment eller trekk. Vi kan også nemne andre venta eigenskapar. Blant anna er fonetiske fenomen drivne av kontekstuelle faktorar som koartikulasjon og taletempo. Fonologiske mønster viser derimot samspel med utprega fonologiske einingar (segment, trykk o. l.); realisasjonen deira er mindre avhengig av den fonetiske konteksten.

I praksis treng vi ofte ganske sofistikerte empiriske data for å skilje mellom fonetiske og fonologiske prosessar. Diverre har vi ikkje så mykje av slikt i litteraturen om samiske språk. Likevel kan vi dra fleire interessante slutningar frå det som vi har av data, som eg prøver å vise nedanfor.

3.4.2. Preaspirasjon som fonetisk regel

Den mest arkaiske formen av preaspirasjon ifølgje livssyklusmodellen er ein fonetisk regel, der dei ustemde klusilane vert realiserte med noko laryngal støy. Denne støyen kan vere lengre eller kortare, men lengda til preaspirasjonen er ikkje sjølvstendig: forskjellige fonologiske strukturar (t. d. korte vs. lange konsonantar) kan vere assosierte med forskjellige eigenskapar til preaspirasjonen, men den har ingen «autonomi» i den fonologiske strukturen. Fonetisk preaspirasjon kan også vere ganske variabel avhengig av faktorar som koartikulasjon og taletempo.

Eitt kriterium som ofte vert brukt for å skilje mellom preaspirasjon som fonetisk regel og eit verkeleg [h]-segment framfor ei klusil er lengd: viss den laryngale støyen er monaleg kortare enn ein vanleg frikativ, er det meir sannsynleg at den representerer ein fonetisk regel, eller, som den tradisjonelle litteraturen ofte skildrar den, ein eigenskap av det påfølgjande segmentet.

I samisk finn vi slik preaspirasjon for det meste i dei austlege språka. Sammallahti (1998: s. 55) nemner her spesielt skoltesamisk og akkalasamisk[16]. Ser vi på materialet åt T. I. Itkonen (2011), finn vi at han bruker symbola <ʾ> («veldig kort stemmelaus vokal») eller <ɔ> («stemmelaus vokal» utan ekstra lengdeteikn) for preaspirasjon i skolte- og akkalasamisk, som nok tyder at preaspirasjon sanneleg er relativt kort i desse varietetane. Vi har også instrumentelle data om skoltesamisk frå McRobbie-Utasi (1991). Eit viktig funn i arbeidet hennar er at lange og korte preaspirerte klusilar viser forskjellig lengd av preaspirasjon, men også av lukkefasen av sjølve klusilen. Det vil seie at preaspirasjonslengd ikkje har noko sjølvstendig fonologisk verd — den følgjer lengda til klusilen, men tek ingen del i den fonologiske strukturen. Eit liknande mønster er også skildra av Ravila (1932) frå den sjøsamiske dialekten i Maattivuono, lengst aust i det nordsamiske området.

Eit spesielt interessant tilfelle finn vi i den umesamiske Nordre Tärna-dialekten. Som vist ovanfor (tabell 7) skilde denne varieteten mellom tre forskjellige typar av preaspirerte klusilar. Det er ikkje godt å seie kva den beste analysen av denne stoda er utan mykje betre data, men den er iallfall ei utfordring for moraisk teori, som til vanleg ikkje tillèt tre gradar av segmentlengd. Ein anonym fagfelle føreslår ei løysing der preaspirasjon i C̀-rekkja er ikkje-moraisk av di den tilhøyrer den neste stavinga ([.ka:$_{\mu\mu}$.htu͡ot]), C̆C-preaspirasjon deler moraen med vokalen ([ka$_{\mu}$[ah·]$_{\mu}$.tu͡on]) og C̀C-preaspirasjon har ein dedikert mora ([ka$_{\mu}$h:$_{\mu}$.tu͡o]). Dette er sjølvsagt mogleg, men inneber ei heller uvanleg plassering av stavingsgrensa i [ka:.htu͡ot]. Eit alternativ er at C̀-preaspirasjon i denne dialekten ikkje hadde gjennomgått stabilisering, og er fortsett ein fonetisk regel. Det vil seie at former [.ka:htu͡ot] *ikkje* har eit fonologisk [h]-segment, men heller ein klusil som får preaspirasjon gjennom ein fonetisk regel. Vi finn ein liknande analyse hjå Bye (2001: s. 132).[17] Viss dette stemmer, viser Nordre Tärna-dialekten eit døme på «regelspreiing» (*rule scattering*),

[16] Denne typen preaspirasjon fanst også og kildinsamisk, t. d. i den utdøydde dialekten i Šoŋguj (E. Itkonen 1971, Sammallahti 1998), som vart tala lengst vest i kildin-området, altså rett ved grensa mot skoltesamisk, men i dag er dette draget nesten borte ifølgje Rießler (2022).

[17] I Iosad (under utarb.) legg eg fram ein analyse av dette slaget, med både fonologisert og stabilisert preaspirasjon, for nokre

der versjonar av det same fenomenet (preaspirasjon) finst i språket på forskjellige stadium av livssyklusen. Vidare er det merkverdig at viss spreiinga av preaspirasjon frå geminatar til sterke klusilar er eit døme av regelgeneralisering, så er stoda i dialekten eit *in vivo*-døme på korleis den innovative generaliserte regelen har gjennomgått fonologisering, men ikkje stabilisering, og dimed prov på at generaliserte reglar går gjennom livssyklusen på akkurat same vis som andre typar lydendringar.

3.4.3. Fonologisk preaspirasjon

I dei fleste samiske språka har preaspirasjon gjennomgått *stabilisering*, og vert best sett på som eit fonologisk segment. Denne slutninga kan vi dra allereie frå tradisjonelle skildringar, der preaspirasjonen vert transkribert som «stemmelaus vokal» med lengdeteikn av typen <ɔ̀> eller <ɔ̄>. Det finst også fonologiske prov på at preaspirasjon har vorte til eit sjølvstendig fonologisk segment. I det meste av nordsamisk kan preaspirasjon vere anten kort eller lang. Denne lengda er uavhengig av lengda på klusilfasen i den påfølgjande konsonanten, men samtidig viser den fonologisk føresett samspel med lengda på den føregåande vokalen (Bals Baal, Odden & Rice 2006) og/eller kvalitet på ein føregåande diftong (Sammallahti 2019). Av grunnar som dette vert den nordsamiske preaspirasjonen analysert som eige segment i den fonologiske overflaterepresentasjonen av Bals Baal, Odden & Rice (2012); i analysen deira vert desse segmenta skipa av ein fonologisk regel frå aspirerte segment, noko som går særs godt saman med livssyklusen. Som vi såg ovanfor viser i det minste lule- og pitesamisk eit liknande system, der preaspirasjonen fungerer som sjølvstendig segment heller enn eit fonetisk drag ved klusilane.

Eit interessant utfall av livssyklusen finn vi i den lulesamiske dialekten i Gällivare. Som skildra av Collinder (1938) viser den konsonantsambanda [xp xt xk] der andre samiske språk har preaspirerte klusilar. Dette kan vi sjå på som eit resultat av at den fonologiske preaspirasjonsregelen vart utvikla til ein ny regel som skipa ein oral heller enn ein laryngal frikativ. Denne oraliseringa er ikkje uvanleg i andre språk med preaspirasjon (Silverman 2003, Clayton 2010), men det spesielle ved akkurat denne utviklinga er at frikativen er [x] uavhengig av artikulasjonsstaden til klusilen. Det normale frå typologisk hald (*pace* Bańczerowski 1969) er at frikativen har same stad som klusilen (type [fp st xk]), men det er altså ikkje det som vi finn i lulesamisk. Det finst ein parallell til denne utviklinga i nokre dialektar av skotsk-gælisk, og Iosad (2020) legg fram at dette mønsteret utviklar seg frå ei lydendring **hk* > *xk* med påfølgjande generalisering av regelen til «h → x framfor alle klusilar». Det er altså mogleg at systemet i Gällivare er eit anna døme på regelgeneralisering.

I aust finn vi stabilisering til frikativar i kildin- og tersamisk ifølgje Rießler (2022). Også her kan vi finne dorsale frikativar [x] eller [ç], avhengig av konteksten, i staden for [h]. Her kunne påverknad frå russisk, som manglar [h] men har både [x] og [xʲ], vere ein faktor.

3.4.4. Særtilfellet enaresamisk

Enaresamisk viser som kjent ei særutvikling. I staden for preaspirerte klusilar viser dette språket *postaspirasjon* ved både lange og korte klusilar. Til forskjell frå preaspirasjon råder det full semje om at postaspirasjon ikkje er noko sjølvstendig segment, men eit drag som tilhøyrer klusilen. Det naturlege scenarioet for utviklinga av den enaresamiske postaspirasjonen er at den kjem frå preaspirasjon, ved at den stemmelause gestusen endrar sitt tilhøve til den orale artikulasjonen (sjå avsnitt 3.1 ovanfor). Denne typen endring er kjend frå spanske dialektar i Andalucia (t. d. O'Neill 2010, Torreira 2012, Ruch & Harrington 2014). Desse dialektane fekk [hC]-samband etter at [s] vart endra til [h] i slutten av ei staving, i ord som *pasta* 'deig'. No kan desse klyngane også vere realiserte som postaspirerte konsonantar ([patʰa]). Sammallahti (1998: s. 55) viser faktisk til nordsamiske dialektar i Aust-Finnmark (altså akkurat mot grensa til skoltesamisk) som kan ha laryngal støy samtidig framfor og etter nokre typar klusilar (det same er mogleg i spansk). Han har sikkert rett i at dette fenomenet er eit slags «bru» mellom (fonetisk) preaspirasjon og det enaresamiske mønsteret.

Som den spanske parallellen viser, er det nok mogleg at enaresamisk postaspirasjon kjem frå eit system

skotsk-gæliske dialektar.

der preaspirasjon allereie har utvikla seg til eit segment, som den har gjort i nordsamisk. Meir generelt legg Sammallahti (1998: s. 193) fram at preaspirasjonen vart til eit sjølvstendig segment allereie i ursamisk, noko som inneber at enaresamisk postaspirasjon kjem frå dette segmentet. Dette er ikkje umogleg, men viss livsyklusmodellen er nokolunde rett, er det ikkje nødvendig. Stabilisering av den fonetiske preaspirasjonen til ein fonologisk regel er noko som vi ventar at kan hende i eit språk der preaspirasjonen har vorte fonologisert, og då er det lett å sjå for seg at stabiliseringa kan skje i kvart språk for seg. I alle fall har vi sjølvstendige utviklingar i vest (nordsamisk og sørover) og aust (kildinsamisk). Med tanke på enaresamisk, er det nok merkverdig at forskjellen mellom korte og lange aspirerte klusilar stod ved lag under endringa frå preaspirasjon til postaspirasjon, og kjem til uttrykk som lengd på klusilfasen: som vi såg ovanfor er det ofte tilfellet (bl. a. i vestleg nordsamisk) at eit fonologisk segment [h] som realisering av preaspirasjon går saman med at klusilfasen har same lengd i Q3 og Q2. Dette gjer det kan hende mindre sannsynleg at det enaresamiske systemet kjem frå eit mønster likt det vi finn i dialektar med stabilisert preaspirasjon[18].

4. Konklusjon

I denne artikkelen har eg lagt fram ei utgreiing om korleis den historiske utviklinga til preaspirasjon i samiske språk kan ha gått for seg, sett frå perspektivet til livssyklusmodellen for lydendring. Modellen har vist seg å vere spesielt produktiv med tanke på både fonetisk og fonologisk variasjon og historisk dialektologi. Dei samiske språka byr på eit rikt laboratorium for å studere språkleg variasjon, og vi har sett at preaspirasjonen kan spele diverse roller i dei fonetiske og fonologiske systema deira. Den synkrone modellen som vi brukte gir fleire gode reiskapar for å forstå skilnadene mellom dei ymse typane preaspirasjon som vi finn i dataa.

Samtidig gir livssyklusmodellen innsikt i den historiske utviklinga av fenomenet. Konsept som fonologisering, stabilisering, og generaliserte reglar gir oss reiskapane vi treng for å spore den historiske traséen, med tanke på både den grammatiske utviklinga og den dialektologiske og geografiske strukturen i dataa.

Til slutt kan vi vende tilbake til spørsmåla som vi stilte i byrjinga. Kva fortel resultata våre om opphavet til preaspirasjon og om rolla til språkkontakt i denne historia? Det finst minst to teoriar som er av interesse her. Ifølgje forfattarar som Posti (1954), Gunnar Ólafur Hansson (2001) og Kortlandt (2019) kunne den samiske preaspirasjonen (eller kan hende meir presist postaspirasjon, som seinare vart til preaspirasjon) vere innlånt frå nordisk. Eitt av argumenta for denne vinklinga er at det finst «mindre» preaspirasjon jo lenger aust vi går i det samiske språkområdet (altså bort frå grensa mot nordiske språk). På den andre sida har forfattarar som Rießler (2004, 2008) og Kusmenko (2008) hevda at den samiske preaspirasjonen kan ha påverka den nordiske.

I rekonstruksjonen som eg la fram i avsnitt 3.3 viser det seg at det geografiske kjerneområdet for preaspirasjon er nordvestsamisk, altså det som er i dag lule- og nordsamisk sone. Preaspirasjon har spreidd seg derfrå mot aust og sør, utan å ha heilt nådd den fjernaste periferien fram til i dag. Eg vil hevde at denne rekonstruksjonen er veldig problematisk for begge typar teoriar om språkkontakt. Med tanke på Posti (1954) sitt argument, skulle vi sikkert ha venta at jo nærmare området kom dei nordiske språka, jo «meir» preaspirasjon bør det ha. Vi kan her godt jamføre lenginga av konsonantar etter korte trykksterke vokalar, som er ganske lik den nordiske kvantitetsomlegginga, og er sanneleg utbreidd spesielt i dei sørlegaste samiske språka. For preaspirasjonen derimot stemmer dette ikkje i det heile teke. Vi har også sett at det finst eit fullgodt scenario for korleis preaspirasjon oppstod i samisk som ikkje krev at vi tyr til språkkontakt.

I seg sjølv byr rekonstruksjonen min på lite som talar for eller imot teorien om samisk opphav for preaspirasjon i nordisk. Vi kunne ikkje peike ut ei eller anna sone som «kjerneområde» for preaspirasjon av geminatar, av di dette fenomenet har spreidd seg til alle språka i greina: dimed er det mogleg at samiske språk allereie hadde det då dei kom i kontakt med det som skulle verte til dagens nordiske varietetar. Svaret om spørsmålet om den samiske preaspirasjonen faktisk spela ei rolle i opphavet av den nordiske er avhengig av andre omsyn, bl. a. i kva grad fonologiske mønster kan verte innlånte, om preaspirasjon er så uvanleg

[18] Merk at austlege dialektar av nordsamisk har ifølgje Sammallahti (2012: s. 162) det same kvantitative mønsteret som skoltesamisk, som i sin tur seiest å vere ei vidareutvikling av det enaresamiske.

på verdsbasis som ein hadde trudd, og tilhøva ved nordisk-samisk språkkontakt; eg gir ei meir inngåande drøfting av alle desse spørsmåla i Iosad (under utarb.).

Takkseiing

Då eg heldt på å fullføre doktorgraden min i Tromsø og var ute på jobbmarknaden, søkte eg ei stilling ved ein norsk kulturinstitusjon. Honoranden sa ja til å skrive eit brev for meg. Etter intervjuet fekk eg høyre frå han at institusjonen ivra etter å vite om eg skulle klare meg i eit miljø der dei faktisk tok nynorsk seriøst. (Eg var på den tida enno bokmålsbrukar.) Svaret frå Trond var: «Han e lingvist. Lingvista kan!» Eg har sidan adoptert dette slagordet som eit slags nordstjerne. Det er altså i denne ånda at eg set fram denne artikkelen som har både samisk og nynorsk i, to emne som eg sterkt assosierer med Trond si mangfaldige verksemd.

Eg takkar også to fagfellar og redaktørane for ekstremt nyttige innspel og rettingar i evalueringsprosessen for denne artikkelen, i tillegg til mykje språkvask; Michael Rießler, som svara på fleire spørsmål under utarbeidinga av artikkelen; og deltakarane på konferansar der eg la fram noko av dette stoffet, spesielt The Fourth Edinburgh Symposium on Historical Phonology (Universitetet i Edinburgh, desember 2019) og 12th International Conference on Nordic and General Linguistics (ICNGL, Universitetet i Oslo, juni 2021). Alle misforståingar og mistak er sjølvsagt mine.

Avstyttingsliste

F = finsk, PSaa = ursamisk, PUr = ururalsk, SaaI = enaresamisk, SaaK = kildinsamisk, SaaL = lulesamisk, SaaN = nordsamisk, SaaP = pitesamisk, SaaSk = skoltesamisk, SaaT = tersamisk. 1 = førsteperson, acc = akkusativ, ess = essiv, gen = genitiv, ill = illativ, inf = infinitiv, loc = lokativ, nom = nominativ, pl = fleirtal, prs = notid, sg = eintal.

Referansar

Aikio, Ante. 2012. An essay on Saami ethnolinguistic prehistory. I Riho Grünthal & Petri Kallio (red.), *A linguistic map of prehistoric Northern Europe* (Mémoires de la Société Finno-Ougrienne 266), 63–117. Helsinki.

Aikio, Ante. 2015. The Finnic «secondary *e*-stems» and Proto-Uralic vocalism. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 95. 25–66. https://doi.org/10.33340/susa.82642.

Bals Baal, Berit Anne, David Odden & Curt Rice. 2006. The phonology of gradation in North Saami. MS., University of Tromsø and The Ohio State University.

Bals Baal, Berit Anne, David Odden & Curt Rice. 2012. An analysis of North Saami gradation. *Phonology* 29(2). 165–212. https://doi.org/10.1017/S0952675712000115.

Bańczerowski, Jerzy. 1969. *Konsonantenalternation im Ostlappischen unter dem Aspekt der Verstärkung-Lenierung: Versuch einer strukturell-phonetischen Analyse* (Prace wydziału filologicznego. Seria Filologia ugrofińska 1). Poznań: Uniwersytet im. Adama Mickiewicza w Poznaniu.

Bartoli, Matteo. 1925. *Introduzione alla neolinguistica: Principi, scopi, metodi* (Biblioteca dell' «Archivum Romanicum». Serie II 12). Genève: Leo S. Olschki.

Bergsland, Knut. 1945. L'alternance consonantique date-t-elle du lapon commun? I *Festskrift til Konrad Nielsen på 70-årsdagen, 28. august 1945* (Studia Septentrionalia 2), 1–53. Oslo: A. W. Brøggers boktrykkeri.

Bergsland, Knut. 1973. Simplification of the Finno-Ugric transcription: Lapp. I Lauri Posti & Terho Itkonen (red.), *FU-transkription yksinkertaistaminen* (Castrenianumin toimitteita 7). Helsinki.

Bermúdez-Otero, Ricardo. 2007. Diachronic phonology. I Paul de Lacy (red.), *The Cambridge Handbook of Phonology*, 497–518. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486371.022.

Bermúdez-Otero, Ricardo. 2015. Amphichronic explanation and the life cycle of phonological processes. I Patrick Honeybone & Joseph C. Salmons (red.), *The Oxford handbook of historical phonology*, 374–399. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199232819.013.014.

Bermúdez-Otero, Ricardo. 2018. Stratal Phonology. I S. J. Hannahs & Anna R. K. Bosch (red.), *The Routledge handbook of phonological theory*, 100–134. London, New York: Routledge. https://doi.org/10.4324/9781315675428-5.

Bermúdez-Otero, Ricardo & Graeme Trousdale. 2012. Cycles and continua: On unidirectionality and gradualness in language change. I Terttu Nevalainen & Elizabeth Closs Traugott (red.), *Handbook on the History of English: Rethinking Approaches to the History of English*, 691–720. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199922765.013.0059.

Blevins, Juliette. 2017. Areal sound patterns: From perceptual magnets to stone soup. I Raymond Hickey (red.), *The Cambridge Handbook of Areal Linguistics*, 55–87. https://doi.org/10.1017/9781107279872.006.

Bull, Tove. 2011. Samisk påverknad på norsk språk. *NOA Norsk som andrespråk* 27. 3–52.

Bybee, Joan & Shelece Easterday. 2019. Consonant strengthening: A crosslinguistic survey and articulatory proposal. *Linguistic Typology* 23(2). 263–302. https://doi.org/10.1515/lingty-2019-0015.

Bye, Patrik. 1997. A generative perspective on «overlength» in Estonian and Saami. I Ilse Lehiste & Jaan Ross (red.), *Estonian prosody: Papers from a symposium*, 36–98. Tallinn: Institute of Estonian Language.

Bye, Patrik. 2001. *Virtual Phonology: Multiple opacity and rule sandwiching in North Saami*. Tromsø: University of Tromsø ph.d.-avh.

Bye, Patrik. 2005. Coda Maximisation in Northwest Saamic. *Nordic Journal of Linguistics* 28(2). 189–221. https://doi.org/10.1017/S0332586505001423.

Bye, Patrik. 2007. Grade alternation in Inari Saami and Abstract Declarative Phonology. I Ida Toivonen & Diane Nelson (red.), *Saami Linguistics*, 53–90.

Bye, Patrik, Ida Toivonen & Elin Sagulin. 2009. Phonetic Duration, Phonological Quantity and Prosodic Structure in Inari Saami. *Phonetica* 66(4). 199–221. https://doi.org/10.1159/000298583.

Clayton, Ian. 2010. *On the natural history of preaspirated stops*. Chapel Hill, NC: University of North Carolina at Chapel Hill ph.d.-avh.

Collinder, Björn. 1929. *Über den finnisch-lappischen Quantitätswechsel: Ein Beitrag zur finnisch-ugrischen Stufenwechsellehre*. Bd. 1: *Einleitung. Ostseefinnisch. Ostlappisch*. Uppsala: Almqvist & Wiksells boktryckeri AB.

Collinder, Björn. 1938. *Lautlehre des waldlappischen Dialektes von Gällivare* (Mémoires de la Société Finno-Ougrienne 74). Helsinki: Suomalais-ugrilainen seura. URN: urn:nbn:fi-fe2016090123410.

Coretta, Stefano. 2020. *Vowel duration and consonant voicing: A production study*. Manchester: University of Manchester ph.d.-avh.

Davis, Garry W. 2008. Toward a Progression Theory of the Old High German Consonant Shift. *Journal of Germanic Linguistics* 20(3). 197–241. https://doi.org/10.1017/s147054270800007x.

Engstrand, Olle. 1987. Preaspiration and the Voicing Contrast in Lule Sami. *Phonetica* 44(2). 103–116. https://doi.org/10.1159/000261784.

Eska, Joseph F. 2018. Laryngeal Realism and the Prehistory of Celtic. *Transactions of the Philological Society* 116(3). 320–331. https://doi.org/10.1111/1467-968x.12122.

Fangel-Gustavson, Nora, Rachid Ridouane & Bruce Morén-Duolljá. 2014. Quantity contrast in Lule Sámi: A three-way system. I *Proceedings of the 10th International Seminar on Speech Processing*, 106–109. Cologne.

Feist, Timothy. 2015. *A grammar of Skolt Saami* (Suomalais-Ugrilaisen Seuran Toimituksia 273). Helsinki: Suomalais-Ugrilainen Seura.

Fertig, David. 2013. *Analogy and morphological change*. Edinburgh: Edinburgh University Press. https://doi.org/10.1515/9780748646234.

Garrett, Andrew & Keith Johnson. 2013. Phonetic bias in sound change. I Alan C. L. Yu (red.), *Origins of sound change: Approaches to phonologization*, 51–97. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199573745.003.0003.

von Gertten, Daniel Z. 2015. *Huvuddrag i umesamisk grammatik*. Oslo: University of Oslo masteroppg.

Gobl, Christer & Ailbhe Ní Chasaide. 1999. Voice source variation in the vowel as a function of consonantal context. I William J. Hardcastle & Nigel Hewlett (red.), *Coarticulation: Theory, data, techniques*, 122–143. https://doi.org/10.1017/CBO9780511486395.006.

Gordon, Matthew. 1997. A fortition-based approach to Balto-Fennic-Sámi consonant gradation. *Folia Linguistica Historica* 18(1–2). 49–79. https://doi.org/10.1515/flih.1997.18.1-2.49.

Gunnar Ólafur Hansson. 2001. Remains of a submerged continent: Preaspiration in the languages of Northwest Europe. I Laurel J. Brinton (red.), *Historical Linguistics 1999: Selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9–13 August 1999* (Current Issues in Linguistic Theory 215), 157–173. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.215.12han.

Hall, Daniel Currie. 2011. Phonological contrast and its phonetic enhancement: Dispersedness without dispersion. *Phonology* 28(1). 1–54. https://doi.org/10.1017/S0952675711000029.

Hejná, Michaela. 2015. *Pre-aspiration in Welsh English: A case study of Aberystwyth*. Manchester: University of Manchester ph.d.-avh.

Helimski, Eugene. 1996. Proto-Uralic gradation: Continuation and traces. I Heikki Leskinen (red.), *Congressus octavus internationalis Fenno-Ugristarum Jyväskylä 10.–15.8.1995*. Bd. 1: *Orationes plenariae et conspectus quinquennales*, 17–51. Jyväskylä: Moderatores.

Hiovain, Katri, Martti T. Vainio & Juraj Šimko. 2020. Dialectal variation of duration patterns in Finnmark North Sámi quantity. *The Journal of the Acoustical Society of America* 147(4). 2817–2828. https://doi.org/10.1121/10.0000994.

Iosad, Pavel. 2020. The life cycle of preaspiration in the Gaelic languages. I Joanna Kopaczyk & Robert McColl Millar (red.), *Language on the move across domains and communities: Selected papers from the 12th triennial Forum for Research on the Languages of Scotland and Ulster*, 200–230. Aberdeen: FRLSU.

Iosad, Pavel. Under utarb. *Phonological drift and language contact: The northern European phonological area*. To appear with Cambridge University Press.

Itkonen, Erkki. 1946. *Struktur und Entwicklung der ostlappischen Quantitätssysteme* (Suomalais-ugrilaisen seuran toimituksia 88). Helsinki: Suomalais-ugrilainen seura.

Itkonen, Erkki. 1971. Ehdotus kildinlapin Šongujn murteen fonemaattiseksi transkriptioksi. I Erkki Itkonen, Terho Itkonen, Mikko Korhonen & Pekka Sammallahti (red.), *Lapin murteiden fonologiaa* (Castrenianumin toimitteita 1), 87–110. Helsinki.

Itkonen, Toivo Immanuel. 1916. *Venäjänlapin konsonanttien astevaihtelu: Koltan, kildinin ja turjan murteiden mukaan* (Mémoires de la Société finno-ougrienne 39). Helsinki: Société finno-ougrienne.

Itkonen, Toivo Immanuel. 2011. *Koltan- ja kuolanlapin sanakirja* (Lexica Societatis Fenno-Ugricae 15). Først publisert 1958. Helsinki: Suomalais-ugrilainen seura.

Kert, Georgiĭ Martynovich. 1971. *Saamskiĭ yazyk (kil'dinskiĭ dialekt): Fonetika, morfologiya, sintaksis*. Leningrad: Nauka.

Kiparsky, Paul. 1988. Phonological change. I Frederick Newmeyer (red.), *Linguistics: The Cambridge survey*. Bd. 1: *Linguistic theory: Foundations*, 363–415.

Kiparsky, Paul. 1995. The phonological basis of sound change. I John Goldsmith (red.), *The handbook of phonological theory*, 640–670. Oxford: Blackwell. https://doi.org/10.1002/9780470756393.ch6.

Kiparsky, Paul. 2008. Fenno-Swedish quantity: Contrast in Stratal OT. I Bert Vaux & Andrew Nevins (red.), *Rules, constraints and phonological phenomena*, 185–220. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199226511.003.0006.

Kiparsky, Paul. 2018. Livonian stød. I Wolfgang Kehrein, Björn Köhnlein, Paul Boersma & Marc van Oostendorp (red.), *Segmental structure and tone*, 195–209. Berlin: Mouton. https://doi.org/10.1515/9783110341263-007.

Korhonen, Mikko. 1981. *Johdatus lapin kielen historiaan* (Suomalaisen Kirjallisuuden Seuran toimituksia 370). Helsinki: Suomalaisen Kirjallisuuden Seura.

Korhonen, Mikko. 1988. The history of the Lapp language. I Denis Sinor (red.), *The Uralic languages: Description, history and foreign influences* (Handbuch der Orientalistik. Achte Abteilung 1), 264–287. Leiden m.m.: E. J. Brill. https://doi.org/10.1163/9789004492493_014.

Korhonen, Mikko, Jouni Mosnikoff & Pekka Sammallahti. 1973. *Koltansaamen opas* (Castrenianumin toimitteita 4). Helsinki.

Kortlandt, Frederik. 2019. On the reconstruction of Proto-Uralic. I Santeri Junttila & Juha Kuokkala (red.), *Petri Kallio rocks: Liber semisaecularis 7.2.2019*, 11–14. Helsinki.

Kristoffersen, Gjert. 2020. Lenisering etter kort vokal: Reliktfenomen eller opphav? *Oslo Studies in Language* 11(2). 225–241. https://doi.org/10.5617/osla.8500.

Kuruch, Rimma Dmitrievna (red.). 1985. *Saamsko-russkiĭ slovar'*. Moscow: Russkiĭ yazȳk.

Kusmenko, Jurij. 2008. *Der samische Einfluss auf die skandinavischen Sprachen: Ein Beitrag zur skandinavischen Sprachgeschichte* (Berliner Beiträge zur Skandinavistik 10). Berlin: Nordeuropa-Institut der Humboldt-Universität zu Berlin.

Kylstra, Andries Dirk. 1967. Zur Substratforschung. *Orbis* 16(1). 101–121.

Kylstra, Andries Dirk. 1972. Die Präaspiration im Westskandinavischen und im Lappischen. *Orbis* 21(2). 367–382.

Lagercrantz, Eliel. 1926. *Sprachlehre des Westlappischen nach der Mundart von Arjeplog* (Mémoires de la Société Finno-Ougrienne 55). Helsinki: Suomalais-ugrilainen seura.

Larsson, Lars-Gunnar. 1990. Glidvokalen i lulesamiskan: En dialektgeografisk undersökning på grundval av Harald Grundströms ordbok. *Svenska landsmål och svenskt folkliv* 113. 167–199.

Larsson, Lars-Gunnar. 2012. *Grenzen und Gruppierungen im Umesamischen*. Wiesbaden: Harrassowitz Verlag.

Lehtiranta, Juhani. 1989. *Yhteissaamelainen sanasto* (Suomalais-ugrilaisen seuran toimituksia 200). Helsinki: Suomalais-ugrilainen seura.

Lehtiranta, Juhani. 1992. *Arjeploginsaamen äänne- ja taivutusopin pääpiirteet* (Suomalais-ugrilaisen seuran toimituksia 212). Helsinki: Suomalais-ugrilainen seura.

Luobbal Sámmol Sámmol Ánte. 2022. Proto-Uralic. I Marianne Bakró-Nagy, Johanna Laakso & Elena Skribnik (red.), *The Oxford guide to the Uralic languages*, 3–27. https://doi.org/10.1093/oso/9780198767664.001.0001.

Luobbal Sámmol Sámmol Ánte & Jussi Ylikoski. 2022. North Saami. I Marianne Bakró-Nagy, Johanna Laakso & Elena Skribnik (red.), *The Oxford guide to the Uralic languages*, 147–177. https://doi.org/10.1093/oso/9780198767664.003.0010.

Magga, Tuomas. 1984. *Duration in the quantity of bisyllabics in the Guovdageaidnu dialect of North Lappish*. Oulu: University of Oulu ph.d.-avh.

McRobbie-Utasi, Zita. 1991. Preaspiration in Skolt Sámi. *SFU Working Papers in Linguistics* 1. 77–87.

Moosberg, Nils Erik. 1920. Stadieväxlingen i Sorsele och Tärna. MS., Uppsala univeristet.

Nickel, Klaus Peter & Pekka Sammallahti. 2011. *Nordsamisk grammatikk*. 2. utg. Karasjok: Davvi girji.

Nielsen, Konrad. 1979. *Lærebok i lappisk (samisk): Utarbeidet på grunnlag av dialektene i Polmak, Karasjok og Kautokeino*. 2. utg. Oslo: Universitetsforlaget.

O'Neill, Paul. 2010. Variación y cambio en las consonantes oclusivas del español de Andalucía. *Estudios de fonética experimental* 19. 11–41.

Odden, David. 1997. Some theoretical issues in Estonian prosody. I Ilse Lehiste & Jaan Ross (red.), *Estonian prosody: Papers from a symposium*, 165–194. Tallinn: Institute of Estonian Language.

Pétur Helgason. 2002. *Preaspiration in the Nordic languages*. Stockholm: Stockholm University ph.d.-avh.

Posti, Lauri. 1953. From Pre-Finnic to late Proto-Finnic: Studies on the development of the consonant system. *Finnisch-ugrische Forschungen* 31. 1–91.
Posti, Lauri. 1954. On the origin of the voiceless vowel in Lapp. *Svenska landsmål och svenskt folkliv* 76–77. 199–209.
Prillop, Külli. 2013. Feet, Syllables, Moras and the Estonian Quantity System. *Linguistica Uralica* 49(1). 1–29. https://doi.org/10.3176/lu.2013.1.01.
Ramsammy, Michael. 2018. The phonology-phonetics interface in constraint-based grammar. I S. J. Hannahs & Anna R. K. Bosch (red.), *The Routledge handbook of phonological theory*, 68–99. London, New York: Routledge. https://doi.org/10.4324/9781315675428-4.
Ravila, Paavo. 1932. *Das Quantitätssystem des seelappischen Dialektes von Maattivuono* (Suomalais-ugrilaisen seuran toimituksia 62). Helsinki: Suomalais-ugrilainen seura.
Ravila, Paavo. 1960. Probleme des Stufenwechsels im Lappischen. *Finnisch-ugrische Forschungen* 33. 285–325. https://doi.org/10.33339/fuf.112711.
Reitan, Jørg. 1930. *Vemdalsmålet: Med oplysninger om andre herjedalske mål*. Oslo: I kommisjon hos Dybwad.
Rießler, Michael. 2004. On the origin of preaspiration in North Germanic. I Karlene Jones-Bley, Angela della Volpe, Martin Huld & Miriam Robbins Dexter (red.), *Proceedings of the Fifteenth Annual UCLA Indo-European Conference* (Journal of Indo-European Studies Monograph 49), 165–185. Washington, D. C.: Institute for the Study of Man.
Rießler, Michael. 2008. Substratsprachen, Sprachbünde und Arealität in Nordeuropa. 54/55. 99–130. https://doi.org/10.1075/nowele.54-55.03rie.
Rießler, Michael. 2022. Kildin Sámi. I Marianne Bakró-Nagy, Johanna Laakso & Elena Skribnik (red.), *The Oxford guide to the Uralic languages*, 219–239. https://doi.org/10.1093/oso/9780198767664.003.0019.
Rießler, Michael & Joshua Wilbur. 2007. Documenting the endangered Kola Saami languages. I Tove Bull, Jurij Kusmenko & Michael Rießler (red.), *Språk og språkforhold i Sápmi* (Berliner Beiträge zu Skandinavistik 11), 39–82. Berlin: Nordeuropa-Institut von Humboldt-Universität zu Berlin.
Ringen, Catherine & Wim A. van Dommelen. 2013. Quantity and laryngeal contrasts in Norwegian. *Journal of Phonetics* 41(6). 479–490. https://doi.org/10.1016/j.wocn.2013.09.001.
Ruch, Hanna & Jonathan Harrington. 2014. Synchronic and diachronic factors in the change from pre-aspiration to post-aspiration in Andalusian Spanish. *Journal of Phonetics* 45. 12–25. https://doi.org/10.1016/j.wocn.2014.02.009.
Salmons, Joseph C. 2020. Germanic laryngeal phonetics and phonology. I Richard B. Page & Michael T. Putnam (red.), *The Cambridge handbook of Germanic linguistics*, 119–142. https://doi.org/10.1017/9781108378291.007.
Sammallahti, Pekka. 1977. *Norjansaamen Itä-Enontekiön murteen äänneoppi* (Suomalais-ugrilaisen seuran toimituksia 160). Helsinki: Suomalais-ugrilainen seura.
Sammallahti, Pekka. 1988. Historical phonology of the Uralic languages, with special reference to Samoyed, Ugric and Permic. I Denis Sinor (red.), *The Uralic languages: Description, history and foreign influences* (Handbuch der Orientalistik. Achte Abteilung 1), 478–554. Leiden m.m.: E. J. Brill. https://doi.org/10.1163/9789004492493_021.
Sammallahti, Pekka. 1998. *The Saami languages: An introduction*. Kárášjohka: Davvi girji.
Sammallahti, Pekka. 2012. On subglottal pulses. I Tiina Hyytiäinen, Lotta Jalava, Janne Saarikivi & Erika Sandman (red.), *Per Urales ad Orientem: Iter polyphonicum multilingue*. Festskrift tillägnad Juha Janhunen på hans sextioårsdag den 12 februari 2012 (Mémoires de la Société finno-ougrienne 264), 359–374. Helsinki.
Sammallahti, Pekka. 2019. *Láidehus sámegiela jietnadatoahpa dutkamii* (Publications of the Giellagas Institute 18). Oulu: Oulu universitehta. URN: urn:isbn:9789526222578.
Sapir, Edward. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace.

Schlachter, Wolfgang. 1958. *Wörterbuch des Waldlappendialekts von Malå und Texte zur Ethnographie* (Lexica Societatis fenno-ugricae 14). Helsinki: Suomalais-ugrilainen seura.
Schlachter, Wolfgang. 1991. *Stufenwechselstörungen in Malålappischen: Aufbau oder Abbau eines Systems?* Wiesbaden: In Kommission bei O. Harrassowitz.
Schuchardt, Hugo. 1885. *Ueber die Lautgesetze: Gegen die Junggrammatiker*. Berlin: Verlag von Robert Oppenheim. URN: urn:nbn:de:kobv:b4-200905195595.
Silverman, Daniel. 2003. On the rarity of pre-aspirated stops. *Journal of Linguistics* 39(3). 575–598. https://doi.org/10.1017/S002222670300210X.
Sjaggo, Ann-Charlotte. 2015. *Pitesamisk grammatik: En jämförande studie med lulesamiska* (Samisk senters skriftserie 20). Tromsø: Septentrio Academic Publishing. https://doi.org/10.7557/10.3591.
Stevens, Kenneth N. & Samuel Jay Keyser. 1989. Primary Features and Their Enhancement in Consonants. *Language* 65(1). 81–106. https://doi.org/10.2307/414843.
Stevens, Kenneth N. & Samuel Jay Keyser. 2010. Quantal theory, enhancement and overlap. *Journal of Phonetics* 38(1). 10–19. https://doi.org/10.1016/j.wocn.2008.10.004.
Stevens, Mary. 2011. Consonant Length in Italian: Gemination, Degemination and Preaspiration. I Scott M. Alvord (red.), *Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*, 21–32. Somerville, MA: Cascadilla Proceedings Project.
Stevens, Mary & John Hajek. 2007. Towards a phonetic conspectus of preaspiration: Acoustic evidence from Sienese Italian. I *Proceedings of ICPhS XVI*, 429–432. Universität des Saarlandes. http://www.icphs2007.de/conference/Papers/1319/1319.pdf (19 juli, 2017).
Stevens, Mary & Ulrich Reubold. 2014. Pre-aspiration, quantity, and sound change. *Laboratory Phonology* 5(4). 455–488. https://doi.org/10.1515/lp-2014-0015.
Suomi, Kari, Juhani Toivanen & Riikka Ylitalo. 2008. *Finnish sound structure: Phonetics, phonology, phonotactics and prosody* (Studia humaniora ouluensia 9). Oulu: University of Oulu.
Tereshkin, Sergeĭ Nikolaevich. 2002. *Yokan'gskiĭ dialekt saamskogo yazȳka*. St Petersburg: Herzen University ph.d.-avh.
Torreira, Francisco. 2012. Investigating the nature of aspirated stops in Western Andalusian Spanish. *Journal of the International Phonetic Association* 42(1). 49–63. https://doi.org/10.1017/s0025100311000491.
Vennemann, Theo. 1972. Phonetic analogy and conceptual analogy. I Theo Vennemann & Terence H. Wilbur (red.), *Schuchardt, the Neogrammarians, and the transformational theory of phonological change*, 181–204. Frankfurt: Athenäum Verlag.
Wagner, Heinrich. 1964. Nordeuropäische Lautgeographie. *Zeitschrift für celtische Philologie* 29(1). 225–298. https://doi.org/10.1515/zcph.1964.29.1.225.
Wiklund, Karl Bernhard. 1896. *Entwurf einer urlappischen Lautlehre*. Bd. 1: *Einleitung, Quantitätsgesetze, Accent, Geschichte der hauptbetonten Vokale* (Mémoires de la Société finno-ougrienne 10.1). Helsingfors: Société finno-ougrienne.
Wilbur, Joshua. 2014. *A grammar of Pite Saami* (Studies in Diversity Linguistics 5). Berlin: Language Science Press. https://doi.org/10.17169/langsci.b17.34.
Wilbur, Joshua (red.). 2016. *Pitesamisk ordbok samt förslag för en pitesamisk ortografi* (Samica 2). Freiburg: Skandinavisches Seminar, Albert-Ludwigs-Universität Freiburg.
Zaĭkov, Petr Mefodievich. 1987. *Babinskiĭ dialekt saamskogo yazȳka: Fonologo-morfologicheskoe issledovanie*. Petrozavodsk: Kareliya.
Zhivlov, Mikhail. 2015. Studies in Uralic vocalism III. *Journal of Language Relationship* 12(1). 113–148. https://doi.org/10.31826/jlr-2015-120109.
Äimä, Frans. 1918. *Phonetik und Lautlehre des Inarilappischen* (Suomalais-ugrilaisen seuran toimituksia 42–43). Helsinki: Suomalais-ugrilainen seura.

Flertalsformer af *ari*-ord i den færøske talesprogsbank[1]

Jógvan í Lon Jacobsen

Fróðskaparsetur Føroya

Abstract

The aim of this article is to investigate dialectal variation of plural endings of *ari*-words in Faroese, i.e., masculine words with *ari*-ending in singular. Such words for example *lærari* 'teacher' may get different plural endings in spoken Faroese: *ar*, *ir*, *a*, and *R* (*lærarar*, *lærarir*, *lærara*, *læraR*[2]). In the written language there is only one correct plural form which is *ar*: *lærarar*. The empirical material in this article is picked up from a corpus of spoken language, *Føroyskur talumálsbanki*, which is a corpus management and analysis system for annotated corpora. The article is also a study of the usability of the corpus concerning dialectal variation in spoken Faroese. The result of the correlation shows that the non-standardized *ir*-variant is most frequent in the corpus. Here I investigate the variation by correlating them with two non-linguistic variables, place, and age.

Keywords: nomina agentis, sociolinguistics, dialectology, Faroese

1. Formål

Formålet med denne artikel er todelt. Dels undersøger jeg variationen af flertalsendelser i ord, som ender på *ari* i ental (ofte omtalt som nomina agentis), fx *lærarar* 'lærere' i ubestemt form i nominativ og akkusativ[3] i forhold til to sociale variabler, alder og sted. Dels vurderer jeg, i hvor høj grad datagrundlaget i talesprogsbanken giver et tilfredsstillende billede af denne variation. Den aldersmæssige variation, som bliver omtalt i artiklen, er eksempel på det, man i sociolingvistikken kalder for tilsyneladende tid, som er en metodisk tilgang, hvor man undersøger sprogændring ved at sammenligne folks tale i forskellige aldre. For så vidt sprogændringer foregår, går man ud fra, at de ældre generationer anvender ældre sprogformer, mens de yngre i højere grad anvender nyere former.

2. Strukturering af artiklen

Før jeg besvarer spørgsmålet om variation i flertalsendelser, er det nødvendigt først at sige noget om det andet spørgsmål: det sproglige materiale i talesprogsbanken. På grund af det vil rækkefølgen blive omvendt: Først omtales talesprogsbanken, og derefter vil variationen blive diskuteret.

Den følgende tekst er opdelt i otte afsnit (afsnit 3 til afsnit 10). Tredje afsnit drejer sig om *ari*- og *i*-afledninger i færøsk, hvorefter der i fjerde afsnit omtales tre centrale begreber i sociolingvistik, nemlig sprogvariation, variabler og varianter. I det femte afsnit behandles projektet, *Ændringer i færøske dialekter*, som ledte frem til oprettelsen af en færøsk talesprogsbank, som omtales i afsnit seks. I syvende afsnit argumenteres der for fordelen ved en standardiseret retskrivning i transskriptionerne. I afsnit otte omtales nogle udfordringer der er knyttet til talesprogsmateriale, hvorefter distributionen af de forskellige flertalsvarianter af *ari*-ord i talesprogsbanken analyseres i niende afsnit. Det tiende og sidste afsnit består af et sammendrag og diskussion om variantfordelingen og datamaterialet i talesprogsbanken.

---

[1] Artiklen bygger på en præsentation i forbindelse med 200 års dagen for V.U. Hammershaimbs fødsel (1819-1909).

[2] *R* = phonetic reduction of unstressed vowel in spontaneus speech.

[3] *ari*-ordene har samme form i nominativ og akkusativ i ubestemt flertal.

 Nordlyd 46.1: 103–113, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigeret af Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publiceret ved UiT Norges arktiske universitet. http://septentrio.uit.no/index.php/nordlyd https://doi.org.10.7557/12.6440

3. *ari-* og *i-endelser*

Nomina agentis refererer til personer, der aktivt udfører eller praktiserer noget, fx *lærari* 'person der underviser andre, især elever i en (folke)skole' og *bakari* 'person der beskæftiger sig erhvervsmæssigt med at bage brød, kager m.m.' (https://ordnet.dk/ddo). Disse ord er afledt af tilsvarende verber (*baka* 'bage' > *bakari* 'bager', *læra* 'lære' > *lærari* 'lærer'). Men mange nomina agentis savner det tilsvarende verb, fx *klokkari* 'klokker'. Indholdet af sådanne afledninger er ikke altid transparent. Eksempelvis er *klokkari* ikke en person, der laver klokker, men en, der bestrider erhvervet at ringe med klokkerne ved en kirke. Nogle nomina agentis er afledt af andre substantiver eller navne, f.eks. *átari* 'grovæder' af substantivet *át* 'spisning' og *B36'ari* 'tilhænger af fodboldklubben *B36*' afledt af navnet på fodboldklubben *B36.* Ingen tilsvarende verber eksisterer for disse ord. Petersen (2019: 79-80) opdeler nomen agentis i tre funktioner:

1) Den gørende (den der udfører en handling): *skrivarin sendi brævið avstað* 'sekretæren sendte brevet afsted'
2) Redskab til at udføre en handling med: *upptrekkjari* 'oplukker', *printari* 'printer'
3) Den oplevende (den der oplever noget): *lurtari* 'lytter'.[4]

Björn Hagström (1977: 47) siger således om nomina agentis i færøsk:

> Många sådana ursprungliga nomina agentis saknar det motsvarande verbet i färöiskan, t. ex. *snikkari*, *skómakari*, *skraddari*. Dermed er förutsätningen given att uppfatta *-ari* som ett suffix för betecknande av yrke utan något verb ligger til grund.

Selvom *ari*-afledningsendelsen er aktiv i moderne færøsk, var der i en periode i 1970'erne en tendens til fra sprogpolitisk side at reducere brugen af denne endelse til fordel for den kortere *i*-endelse, muligvis inspireret af islandsk (jf. fx isl. *blóðgjafi*, fær. *blóðgevi* 'bloddonor' og isl. *sæðisgjafi*, fær. *sáðgevi* 'sæddonor', se islex.arnastofnun.is). i-endelsen er meget tydelig i en liste med geografiske navne med tilsvarende inkolentnavne, udgivet af de nordiske sprognævn i 1974 (jf. *Navne på stater* i litteraturlisten), fx *afgani* 'afganer' i stedet for *afganari* og *amerikani* 'amerikaner' i stedet for *amerikanari*. I oversigten over inkolentnavne på Sprogrådets hjemmeside er denne praksis ændret, således at mange af disse inkolentnavne nu har dobbeltformer, fx *afganari/afgani* og *amerikanari/ amerikani* (malrad.fo/orðalistar/lond og tjóðir, 4. januar 2022).

Denne orddannelsesmåde med *i*-endelser i stedet for *ari* er i øvrigt velkendt i dannelsen af afløsningsord i færøsk. En optælling i en nyordsliste viser, at af tyve afløsningsord, som ender på *ari* og *i*, er der ni med *ari-* og elleve med *i*-afledning (Poulsen, 2004: 519-524). Eksempler på *ari* er: *ambætari* 'server', *atstøðari* 'assistent' og *blekkprentari* 'blækprinter'. Og af eksempler på *i* kan nævnes *ostskeri* 'ostehøvl', *floksherji* 'partisan', *hjáseti* 'bisidder, *sjókagi* 'søkikkert', *skyni* 'detektor', *vevkagi* 'browser' og *vísi* 'cursor'.

Weyhe (2012a; 2012b; 2012c; 2015: 423) drøfter geografisk variation af flertalsformer af nomina agentis historisk og konkluderer, at *ir*-endelsen breder sig (Denne undersøgelse bekræfter rigtigheden i hans konklusion, især hvis man lægger *ir* og *R* sammen (jf. afsnit 9 nedenfor). Hagström (1977: 47-52) siger, at danske importord, som ender på *er* og *or* får *ari*-endelse i færøsk. Her kan jeg tilføje, at der også findes eksempler på, at danske importord med udlydende *or*, kan få *ur*-endelse, fx *motorur* (da. motor), *traktorur* (da. traktor), *radiatorur* (da. radiator). På den anden side er der også eksempler på, at ord, som refererer til personer, og som på dansk ender på *or*, får tilføjet *ur* i færøsk, fx *senatorur* 'senator' (mask.) og *faktorur* 'opsynsmand, forvalter'[5] (mask.), ikke **senatorari*, **faktorari* (jf. fx *doktari*, *professari*).

[4] Her drejer det sig kun om egentlige nomina agentis, og derfor er den semantiske kategori indbygger ikke medregnet, fx *kanadiari* 'kanadier'.

[5] Om betydningen af faktor se *faktor* 1.2. i *Ordbog over det danske sprog*, https://ordnet.dk/ods, 4. januar 2022.

4. Sprogvariation, variabler og varianter

Da denne artikel er en deskriptiv sociolingvistisk undersøgelse af sprogvariation i færøsk talesprog, vil jeg kort berøre nogle centrale træk ved sociolingvistik. Formålet med sociolingvistiske variationsundersøgelser er at opnå større bevidsthed om sprogbrug i social sammenhæng. Det betyder, at foruden at undersøge en variants geografiske distribution, relateres sprogbrugen også til sprogeksterne (sociale) variabler. Chambers og Trudgill (1990: 54) siger, at alle dialekter både er lokale og sociale: "All dialects are both regional and social, since all speakers have a social background as well as a regional location." Hudson (1999: 3) siger om forskellen mellem sociolingvistik og lingvistik, "that linguistics differs from sociolinguistics in taking account only of the *structure* of language, to the exclusion of the social contexts in which it is learned and used." En sociolingvistisk dialektundersøgelse siger derfor ikke blot noget om geografisk distribution (som traditionelle dialektundersøgelser gør), men afdækker samtidig i hvilken grad distributionen af en sproglig variabel kan knyttes an til sociale variabler, i dette tilfælde alder og sted.

I sociolingvistiske variationsundersøgelser arbejdes der med variabler og varianter. En variabel er en størrelse, som varierer, og varianterne er de forskellige former eller udtryk, som variablen kan realiseres som. Den sproglige (sproginterne) variabel i denne undersøgelse er flertalsendelser af *ari*-ord (fx *lærarar* 'lærere') med varianterne *lærarar*, *lærarir*, *lærara* og *læraR*. Den sociale (sprogeksterne) variabel er en størrelse, der ligger uden for selve sproget, fx alder, geografisk sted, køn osv.

En sociolingvistisk korrelationsstudie gør det muligt at undersøge, i hvilken grad der er sammenhæng mellem de sproglige og sociale variabler. Af pladshensyn har jeg begrænset mig til at undersøge betydningen af to sociale variabler, nemlig alder og sted. Stedsvariablen siger noget om distribution og frekvens af varianterne i de forskellige dialekter, og aldersvariablen afdækker, i hvor høj grad der er sammenhæng mellem distribution af de fire flertalsformer og informanternes alder. Aldersvariablen er naturligvis interessant i en undersøgelse i tilsyneladende tid. I analysen bliver den sproglige variabel korreleret med de sociale variabler for at undersøge, om der er nogen sammenhæng. Formålet er at afdække eventuelle mønstre i variationen. Når vi taler om mønstre i sociolingvistisk forstand, hentyder det til, at variationen af en sproglig variabel knyttes an til en social variabel og kan forklares ud fra den. For exempel korrelerer udtalevariation ofte med alder.

Traditionelle dialektundersøgelser viser geografisk variation, fx at folk, som bor syd for Skopunarfjørður, siger [gaŋga] 'gå', mens folk nord for Skopunarfjørður siger [geŋga] (se kort 1). Dette faktum er naturligvis interessant, set i relation til geografien, men det siger intet om, i hvilket omfang sproglig variation forekommer internt i det pågældende dialektområde. Netop sproglig variation er tema i denne artikel, hvor der fokuseres på interindividuel variation i forhold til alder og geografi.

5. Ændringer i færøske dialekter

I 2015 bevilgede Granskingarráð Føroya (det færøske forskningsråd) midler til et treårigt projekt, hvis formål var at undersøge ændringer i færøske dialekter. Projektet blev lavet i samarbejde mellem færøske og norske sprogforskere og var stedfæstet på Føroyamálsdeildin ved Fróðskaparsetur Føroya. Deltagerne i projektet var Hjalmar P. Petersen, Jógvan í Lon Jacobsen og Heðin Jákupsson, Fróðskaparsetur Føroya, og Helge Sandøy, Universitetet i Bergen, og Edit Bugge, Høgskulen på Vestlandet. Paul Meurer, sprogteknolog ved Universitetet i Bergen, havde udviklet programvaren Corpuscle, som den norske *Talebanken* var udformet i, og han oprettede også den færøske *Føroyskur talumálsbanki* som et parallelt korpus og lagde det færøske materiale ind i den (jf. afsnit 6 nedenfor). Trond Trosterud, professor i sprogteknologi ved Universitetet i Tromsø, gjorde en grammatisk tagger, som er knyttet til talesprogsbanken. Studerende og ansatte ved instituttet hjalp til med feltarbejdet.

Projektet, som denne artikel bygger på, havde titlen *Ændringer i færøske dialekter gennem to generationer – hvordan, hvor hurtigt og hvorfor?* Der var tale om en sociolingvistisk studie med det formål at få indsigt i relationen mellem samfundsændringer og sprogændringer. For at få den vinkel med var projektet bygget op omkring modsætningen centrum-periferi, og interviewpersonerne var udvalgt efter denne model. Vi udvalgte fem steder spredt over hele landet: Tórshavn, Vágar, Eysturoy (Eiði), Norðoyggjar (Klaksvík, Viðareiði) og Suðuroy (se kort 1 nedenfor). Ifølge dette koncept var Tórshavn tænkt som centrum i forhold til Vágar og Eiði, mens Klaksvík var tænkt som centrum i forhold til Viðareiði,

og Tvøroyri som centrum i forhold til andre bygder på Suðuroy. Denne model er inspireret fra projektet *Dialektendringsprosessar* ved Universitetet i Bergen, og vi syntes, det kunne være interessant at undersøge, om vi kunne se de samme tendenser, som man ser i Norge, nemlig at de såkaldte regionscentre påvirker de mindre dialektområder, som ligger tæt på disse centre (jf. Sandøy, 2008 og Sandøy, 2017). Dette spørgsmål drøftes ikke i denne artikel.

Kort 1: Kort over Færøerne (Kilde: Wikimedia Commons)

Det er naturligvis nødvendigt med en bred repræsentation af informanter, når sociolingvistiske studier skal udføres. Har man en fornuftig spredning med hensyn til alder, geografi, køn osv. kan man få interessante oplysninger om sproglig praksis i sociolingvistisk perspektiv.

Informanterne bestod af tre aldersgrupper. De yngste informanter var 15-årige elever, som gik i 9. klasse. I den mellemste aldersgruppe var informanterne 45-65 år gamle, og i den ældste aldersgruppe var informanterne 70+. Den oprindelige plan var at have mindst fire personer i hver aldersgruppe, dvs. 12 informanter fra hvert dialektområde. Men antallet ligger langt over det, i særdeleshed i Tórshavn, som har et samlet antal informanter på 37. Foruden nyindsamlet materiale brugte vi nogle gamle optagelser. Testdesignet i hele projektet bestod af tre metoder: Interviewer, spørgeskemaer om holdning til dialekter og masketester. Denne artikel bygger kun på interviewene[6].

Der er i alt 69 interviewer med 103 informanter, 54 kvinder og 49 mænd. Årsagen til at antallet informanter er så stort i forhold til interviewene er, at i de fleste tilfælde sad to og to informanter og talte sammen. Denne metode fungerede godt i de fleste tilfælde, da informanterne kendte hinanden i forvejen.

6. Talesprogsbanken

Alle optagelserne er lagt ind i talesprogsbanken som lydfiler sammen med transskriptioner. 30 minutter af hver samtale blev transskriberet med standardretskrivning. Transskriberingen blev gjort i Praat-

[6] Mere kan læses om de færøske holdningsundersøgelser i Bugge (2018) og í Bugge og Jacobsen (2018).

programmet, som egner sig godt til talesprogsdata. Transskriberingen var færdig ved udgangen af 2018. Det samlede antal ord i talesprogsbanken pr. januar 2022 er 471 178.

Når vi nu har adgang til talesprogsmateriale, giver det nye muligheder for at forske i sprogvariation. Talesprogsbanken er et godt arbejdsredskab til sociolingvistiske undersøgelser. Der er mange tekniske funktioner indbygget i programmet, som gør det muligt at lave korrelationsstudier. Den færøske talesprogsbank er en del af den norske talesprogsbank, Clarino (Common Language Resources Infrastructure Norway), som er udviklet ved Universitetet i Bergen.

Alle ord i talesprogsbanken er tagget med grammatiske oplysninger, køn og tal for substantiver, præsens og præteritum for verber osv. Ud over grammatiske oplysninger er der adgang til metaoplysninger om informanterne, fx fødselsår, hjemsted, køn osv. I tabel 1 ser vi et eksempel på tagging og metasproglige oplysninger med et eksempel på skriftsprogsformen *lærarar* (som i øvrigt forekommer 53 gange i talesprogsbanken):

word:	Lærarar
lemma:	Lærari
pos:	N
features:	N Msc Pl Nom Indef
person:	30716
document:	KKG_2TO
age:	G
is-interviewer:	Nei
dialect:	Tórshavn
sex:	K
birthyear:	1945
place:	Tórshavn
admission-year:	2016
interviewer:	90215
admission-place:	Tórshavn
mother-from:	Hvalba
father-from:	Svínoy

Tabel 1: Eksempel på metaoplysninger i talesprogsbanken.

7. Standardretskrivning

I transskriptionerne anvendes standardretskrivning og standardbøjning. I udgangspunktet skal transskriptøren skrive det, som bliver sagt, dvs. at transskriptionen skal være en deskriptiv og objektiv gengivelse af talen. Transskriptøren skal koncentrere sig om at forstå og gengive det, som høres på lydfilen uden samtidig at tænke på fonetiske forskelle. I øvrigt er transskriptøren sjældent kendt med alle de forskellige variabler og varianter, som forskere på et senere tidspunkt måtte være interesseret i. Det kan desuden være yderst problematisk at identificere en bestemt udtale af et ord, og den tolkning skal transskriptøren ikke foretage. Derfor valgte vi at transskribere dialektale varianter, fx *lærarir* og *lærara* (flertal af *lærari* 'lærer') med standardformen *lærarar*. Et andet eksempel på standardisering i transskriptionen var, at den dialektale flertalsform *hevði* 'havde' (præteritum af verbet *hava* 'have') blev skrevet *høvdu*, som er den korrekte form ifølge standardgrammatikken. En søgning på formen *høvdu* i søgestrengen giver derfor samtlige belæg på begge udtalevarianter i præteritum flertal. Det betyder dog ikke, at denne transskriberingspraksis er den eneste rigtige eller mest hensigtsmæssige, fordi når man har adgang til en grammatiktagger, kan man nemt lave en avanceret søgning på grammatiske kriterier og opnå de samme søgeresultater. Den største fordel med standardretskrivning er i de tilfælde, hvor der eksisterer mange talesproglige varianter af et enkelt ord. I mange tilfælde forekommer disse talesprogsvarianter aldrig på skrift, og derfor kan det være vanskeligt at gengive sådanne ord med en lydret ortografi. Som eksempel på det vil jeg anføre adverbiet *soleiðis* [so:lai:jɪs] 'således', som kan have mange forskellige udtaler (i eksemplerne her har jeg bibeholdt *ð*, selvom

det ikke udtales), fx *soloyðis*, *soloyðus*, *soloyðs*, *soleiðsnar*, *soleiðsna*, *soleiðsni*, *soloyðsni*, *soloyðsnar*, *soloyðsna, soeiðsna, soeiðsnar* osv. Dette eksempel skulle gerne vise fordelen ved at anvende standard-retskrivning. At skrive en tilnærmelsesvis lydret gengivelse af udtalen af hvert enkelt ord ville være en umulig opgave. Ved at skrive standardformen ind i søgestrengen får man alle udtaler af vedkommende ord.

8. Udfordringer med talesprogsmateriale

Det er spændende men samtidig udfordrende at arbejde med talesprogsmateriale. Det er spændende på den måde, at vi nu for første gang har mulighed for at arbejde med et annoteret korpus med talesprogsmateriale. Udfordringen er bl.a., at andre normer gør sig gældende i talesproget end i skriftsproget. En tilbagevendende udfordring med talesprog er, at udtalen ikke altid er så distinkt, og det går især ud over endelserne, som kan være meget svagt artikuleret og til tider kan være helt væk. Når to personer sidder og taler sammen, som de gjorde i denne undersøgelse, afbryder de ofte hinanden eller taler i munden på hinanden. I enkelte tilfælde afbryder intervieweren samtalen. Ofte er der højlydt latter. Lydkvaliteten på optagelsen kan være dårlig. Der kan være støj i baggrunden. Somme tider bliver informanterne så ivrige, når de fortæller en spændende historie, at taletempoet stiger voldsomt. En spændende historie kan således medføre udistinkt udtale. Nogle personer har en utydelig udtale, andre taler hurtigt osv. Alle disse forhold kan gøre det vanskeligt at høre, hvordan informanterne rent faktisk udtaler ordene. Derfor er der mange tvivlsspørgsmål, når endelserne skal undersøges. Det kommer også tydeligt til udtryk i denne undersøgelse, hvor det kan være vanskeligt at holde *ir*-endelsen og *R*-endelsen ude fra hinanden.

9. *ari*-ord i talesprogsbanken

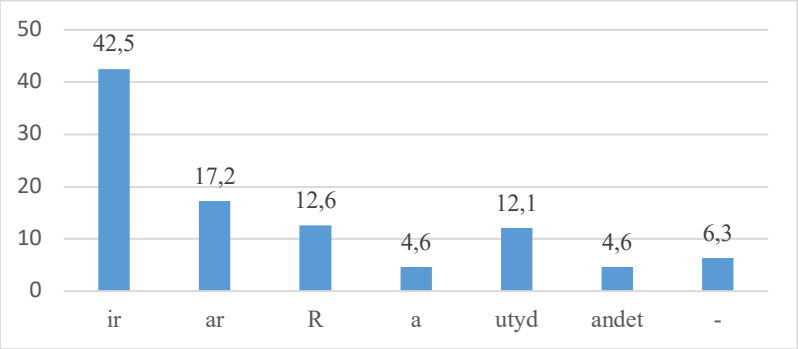
Nu vil jeg undersøge distributionen af de forskellige varianter af *arar*-endelsen i talesprogsbanken. En søgning på skriftsprogsendelsen *arar* (skrevet ".*arar" i søgestrengen) giver 174 belæg eller forekomster, som opfylder betingelsen for ord, der ender på *arar*. Det burde være tilstrækkeligt til en dialektal variationsundersøgelse. Men det blev hurtigt klart for mig, at der var en del udfordringer i materialet. En fejl i en af lydfilerne gjorde, at lyden ikke kunne høres, og derfor er disse eksempler ikke annoteret. Senere er den blevet lagt ind på ny. I denne lydfil er der 19 belæg på *arar*. Tilbage var der så 155 belæg. Af disse 155 belæg måtte 21 grupperes for sig på grund af utydelig udtale (markeret som *utyd*). Tilbage var der nu 134 belæg, som jeg havde til rådighed i min analyse. Men yderligere to belæg måtte fjernes, fordi de blev brugt af intervieweren. Derfor består analysen af 132 belæg (jf. tabel 2). Når jeg gennemlyttede alle belæggene, blev det klart for mig, at nogle personer brugte en endelse, som hverken var *ar*, *ir* ella *a*. Her kunne jeg rent faktisk ikke høre nogen endelse i det hele taget. Den tryksvage vokal var bortsynkoperet. Denne udtale mener jeg er eksempel på stavelsesbærende *r* (skrevet *R*), fx *danskaR*, *skutaR.* Sandøy (1994: 15) har undersøgt stavelsesbærende *n* i færøsk og siger, at den lyd, som falder ud, enten er *i* eller *u* og at det rammer den sidste stavelse i ordet. I dette tilfælde er det også den sidste stavelse, som bliver bort-synkoperet, fx *skutarir* > *skutaR*. Jeg er enig med den ene fagfællebedømmer, som siger, at hvis *r'et* er stavelsesbærende, skal man have en fornemmelse af, at rytmen forudsætter tre stavelser.

Det kan være vanskeligt at høre forskel på *R* og *ir*, især når folk taler hurtigt. Siger de fx *danskaR* eller *danskarir*? Og jeg skal erkende, at jeg ofte var i tvivl og stadigvæk er det, men ved at lytte igen og igen er jeg kommet frem til 22 belæg (12,6%) på *R* (stavelsesbærende *r*). Hvis jeg skulle klassificere *R*-formerne i en af de andre grupper, ville de komme i *ir*-gruppen, fordi *R*-formerne fonetisk ligger nærmere *ir* end *ar* (*a*-kategorien er udelukket). En sådan artikulatorisk forklaring er vel ikke utænkelig. Afstanden mellem gane og tunge er meget lille ved de høje vokaler *i* og *u*, og det samme gør sig gældende ved udtalen af konsonanten *r*. Derved kan der opstå en slags "assimilation".

En foreløbig akustisk undersøgelse af de forskellige endelsers (suffiksers) varighed viser, at *R*-varianten i gennemsnit har en kortere varighed end *ir*- og *ar*-endelserne, hhv. 155 ms for *R* og 282 ms for *-ir* og *-ar*. Resultaterne skal ses med forbehold grundet lydens kvalitet og analysens foreløbige grovhed. Men målingerne viser en gennemgående tendens til, at R-varianterne udtales med kortere varighed end *ir*- og *ar*-varianterne, hvilket underbygger argumentet for kategoriseringen[7].

[7] Tak til Iben Nyholm Debess for den akustiske undersøgelse, som blev lavet i Praat.

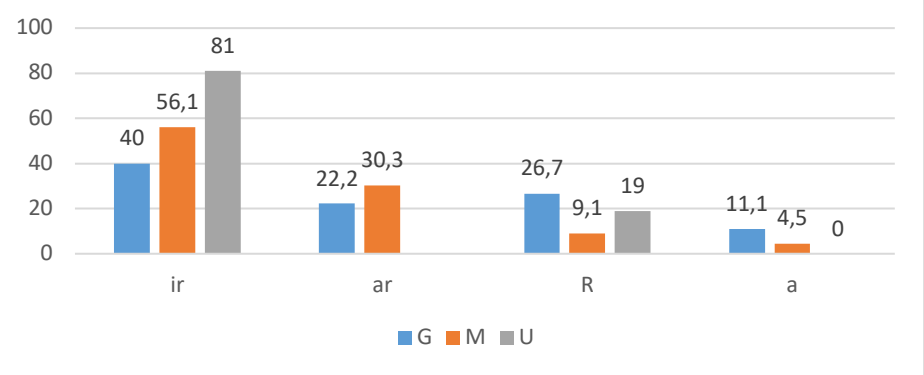
Søjlediagrammet i figur 1 viser den procentuelle distribution af varianterne (medregnet de belæg, der falder uden for varianterne). I tabel 2 nedenfor anføres antallet af belæg sammen med procenttallene fordelt på de tre aldersgrupper.



Figur 1: Den procentuelle distribution af varianterne

I gruppen *andet* placeres fx verber, som ender på *arar*, eksempelvis *svarar* 'svarer'. I gruppen længst til højre med en streg under placeres de uannoterede belæg (pga. en fejl i en lydfil, se ovenfor). Forkortelsen *utyd* betyder utydelig udtale. Distributionen viser meget tydeligt, at *ir*-varianten er den mest frekvente, og hvis man lægger *ir*- og *R*-varianten sammen (hvilket jeg er tilbøjelig til), får man et resultat, som viser, at over halvdelen af samtlige belæg falder ind under *ir*-varianten.

Figur 2 er en visualisering af tabel 2, som viser den procentuelle fordeling i forhold til alder:



Figur 2: Den procentuelle distribution af varianterne i forhold til alder.

	sum	**ir**	**ar**	**R**	**a**
sum	**134** (100,0)	**74** (55,2)	**30** (22,4)	**22** (16,4)	**8** (6,0)
	134 (100,0)	74 (69,3)	30 (13,1)	22 (13,7)	8 (3,9)
	2 (100,0)	2 (100,0)			
G	45 (100,0)	18 (40,0)	10 (22,2)	12 (26,7)	5 (11,1)
M	66 (100,0)	37 (56,1)	20 (30,3)	6 (9,1)	3 (4,5)
U	21 (100,0)	17 (81,0)		4 (19,0)	

Tabel 2: Den procentuelle distribution af varianterne i forhold til alder

Tallene i tabel 2 er antal belæg og ikke antal informanter. Distributionen viser, at *ir*-varianten er den hyppigst forekommende i alle aldersgrupper. De yngste bruger udelukkende *ir*- og *R*-varianten. Den høje

frekvens af *ir*-formen betyder vel egentlig, at den har fået status som en overdialektal form, brugt af folk overalt i samfundet.

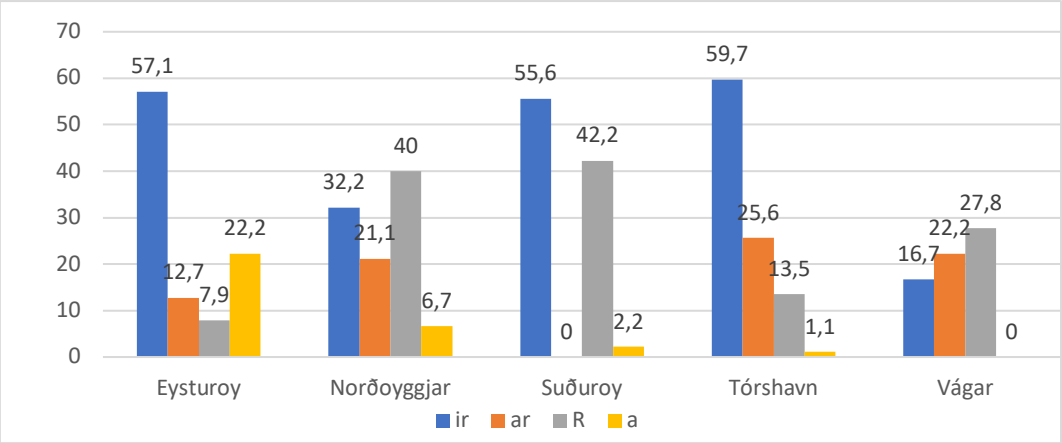
Mens tabel 2 viser distributionen af belæg, viser tabel 3 fordelingen af informanter mellem de forskellige dialektområder. Tabel 3 viser, hvordan de forskellige dialektområder er repræsenteret i forhold til alder og køn:

		Eysturoy	Suðurstreymoy	Norðoyggjar		Suðuroy		I alt
Alder	**Vágar**	**Eiði**	**Tórshavn**	**Klaksvík**	**Viðareiði**	**Tvøroyri**	**Andre bygder**	
G	2k+2m	2k+2m	7k+6m	2k+2m	2k+2m	2k+3m	2k+1m	37
M	2k+2m	2k+2m	6k+6m	1k+1m	2k	1k+2m	3k	30
U	2k+2m	2k+2m	6k+6m	2k+2m	2k+2m	2k+2m	2k+2m	36
I alt	**12**	**12**	**37**	**10**	**10**	**12**	**10**	**103**

Tabel 3: Antal informanter fordelt på alder og køn i de undersøgte dialektområder (k=kvinder; m=mænd)

Kolonnen yderst til højre viser, at antallet informanter i de tre aldersgrupper ligger på nogenlunde samme niveau, mellem 30 og 37. Kønsfordelingen er også rimelig (54 kvinder og 49 mænd). Til gengæld er antallet af informanter fra Tórshavn langt større end fra de andre steder. Årsagen til det forholdsvis store antal er, at vi i et underprojekt ville få mulighed for at undersøge, hvorvidt tilflytning til Tórshavn påvirker folks sproglige praksis. Derfor blev informanterne fra Tórshavn udvalgt efter andre kriterier og opdelt i tre undergrupper: (I) Informanten selv var tilflytter, (II) Informanten havde to tilflytterforældre, og (III) Begge informantens forældre var fra Tórshavn.

Figur 3 viser variantdistributionen mellem de fem dialektområder:



Figur 3: Den procentuelle distribution af varianter i de fem dialektområder

Diagrammet viser, at frekvensen af *ir*-formerne er særlig høj på Eysturoy, Suðuroy og i Tórshavn. På den anden side ligger skriftformen *ar* relativt lavt. Tabel 4 viser en distributionsanalyse, hvor jeg krydser sprogvariant, alder og sted. Tallene i parentes er procenttal af forekomster i hele gruppen, og summen af dem giver 100 procent på den vandrette akse. På den vandrette akse øverst i tabellen står de sproginterne varianter og på den lodrette akse i venstre side står de sprogeksterne variabler, sted og alder. Øverst i den første talkolonne står det samlede antal belæg på den undersøgte variabel og nedenfor står de samlede belæg for hvert dialektområde, fordelt på de tre aldersgrupper. Derefter kommer belæggene på hver af de fire varianter: *ir*, *ar*, *R* og *a.* (rest) som står øverst til venstre i tabel 4 er to eksempler på spørgsmål fra interviewveren.

		Sum	ir	ar	R	a
Sum		**134** (100,0)	**74** (55,2)	**30** (22,4)	**22** (16,4)	**8** (6,0)
(rest)		2 (100,0)	2 (100,0)			
		2 (100,0)	2 (100,0)			
Eysturoy		**22 (100,0)**	**12 (57,1)**	**4 (12,7)**	**2 (7,9)**	**4 (22,2)**
	G	6 (100,0)		1 (16,7)	1 (16,7)	4 (66,7)
	M	14 (100,0)	10 (71,4)	3 (21,4)	1 (7,1)	
	U	2 (100,0)	2 (100,0)			
Norðoyggjar		**15 (100,0)**	**5 (32,2)**	**4 (21,1)**	**4 (40,0)**	**2 (6,7)**
	G	3 (100,0)	2 (66,7)	1 (33,3)		
	M	10 (100,0)	3 (30,0)	3 (30,0)	2 (20,0)	2 (20,0)
	U	2 (100,0)			2 (100,0)	
Suðuroy		**21 (100,0)**	**15 (55,6)**		**5 (42,2)**	**1 (2,2)**
	G	15 (100,0)	10 (66,7)		4 (26,7)	1 (6,7)
	M	5 (100,0)	5 (100,0)			
	U	1 (100,0)			1 (100,0)	
Tórshavn		**65 (100,0)**	**38 (59,7)**	**18 (25,6)**	**8 (13,5)**	**1 (1,1)**
	G	18 (100,0)	5 (27,8)	8 (44,4)	5 (27,8)	
	M	31 (100,0)	18 (58,1)	10 (32,3)	2 (6,5)	1 (3,2)
	U	16 (100,0)	15 (93,7)		1 (6,2)	
Vágar		**9 (100,0)**	**2 (16,7)**	**4 (22,2)**	**3 (27,8)**	
	G	3 (100,0)	1 (33,3)		2 (66,7)	
	M	6 (100,0)	1 (16,7)	4 (66,7)	1 (16,7)	
	U	0 (0,0)				

Tabel 4: Antal belæg af de fire varianter i forhold til alder og sted.
G = gamle; M = midaldrende; U = unge.

Som det fremgår af tabel 4, er der 74 belæg på *ir*-varianten (svarende til 55,2% af samtlige belæg). Dermed er *ir*-varianten den mest frekvente i hele korpusset. Til sammen udgør *ir*- og *R*-varianten 71,6%. Det er ikke usandsynligt, at *ir*-varianten vil brede sig endnu mere i fremtiden som en slags statusform (jf. Weyhe, 2012b: 368). Den næststørste kategori er *ar*-varianten (skriftsprogsvarianten), hvis brugsfrekvens dog ligger meget lavere end *ir*-varianten med 30 belæg (svarende til 22,4%). Mindst frekvent er *a*-varianten med otte belæg (svarende til 6,0%). Af de otte *a*-former, findes fire hos gamle mennesker på Eysturoy og to hos midaldrende fra Norðoyggjar. Der er altså ingen forekomst af *a*-former hos de yngste. Den før så udbredte *a*-form på Eysturoy og i Norðoyggjar er stærkt på retur og afløses af *ir*-former, som i øvrigt forekommer i alle aldersgrupper over hele landet. *R*-formerne forekommer i alle aldersgrupper, dog med en lille overvægt i Tórshavn og på Suðuroy[8].

10. Sammendrag og diskussion

Formålet med denne artikel er dels at undersøge en morfologisk variabel og dens manifestationer i talesproget, dels at undersøge anvendeligheden af talesprogsbanken som sprogteknologisk værktøj til sociolingvistiske variantionsundersøgelser.

Den undersøgte variabel er flertalsformer af *ari*-ord i nominativ og akkusativ i ubestemt form. Skriftsprogsvarianten er *ar*, mens talesproget desuden har varianterne *ir*, *a* og *R* (fx *lærarar*, *lærarir*, *lærara*, *læraR*). Undersøgelsen viser meget tydeligt, at *ir*-varianten er den mest frekvente i hele materialet med 74 belæg (55,2%), derefter kommer skriftsprogsvarianten *ar* med 30 belæg (22,4%), *R*-varianten med 22 belæg (16,4%) og *a*-varianten med otte belæg (6,0%).

[8] Suðuroy-varianten med en *o*- eller *ø*-lignende lyd tolker jeg som variant af *i*.

I den yngste aldersgruppe på Eysturoy er der kun to belæg på den undersøgte variabel, og på Suðuroy er der kun et enkelt belæg i den yngste aldersgruppe. På Vágar er der intet belæg blandt de yngste af et samlet antal på ni belæg på Vágar. Det siger sig selv, at når man nedbryder ni belæg på tre aldersgrupper, bliver tallene meget små og usikre. Datagrundlaget er yderst spinkelt, hvilket påvirker reliabiliteten i negativ retning og gør konklusionerne uvisse. På den anden side er datamængden for Tórshavn betydeligt større, hvilket øger undersøgelsens reliabilitet. Derfor er det nødvendigt at udvide talesprogsbanken med nyt materiale, især fra områderne uden for Tórshavn for på den måde at få et mere balanceret korpus. Datagrundlaget må altså øges betragteligt, før vi kan få fuldt udbytte af alle de tekniske muligheder, der ligger i analyseprogrammet. Trods den begrænsede datamængde viser materialet nogle tendenser i variationen i forhold til alder og sted.

Undersøgelsen viser, at *ir*-varianten står stærkt blandt de unge. Det er derfor ikke usandsynligt, at denne variant vil brede sig endnu mere i fremtiden som en overdialektal form (jf. Weyhe, 2012b: 368).

Distributionen med *ir*-varianten i top åbner op for en helt anden diskussion, nemlig hvorvidt tiden ikke er inde til at acceptere *ir*-formen som tilladt skriftsprogsform ved siden af *ar*-formen. Det kan virke lidt besynderligt, at den mest frekvente talesprogsvariant ikke er tilladt i skrift. Den afgørelse ligger hos Málráðið (Sprogrådet).

Erfaringerne fra dette studie er, at rent teknisk er talesprogsbanken et fortræffeligt redskab til sociolingvistiske dialektstudier og variationsstudier i øvrigt. Nu er det muligt at udføre kvantitative dialekt- og dialektforandringsstudier med udgangspunkt i et dialektkorpus, som er designet til netop dette formål.

Ulempen ved det færøske dialektkorpus er dens begrænsede størrelse med knap en halv million ord. Når man undersøger en variabel med alder og sted, bliver tallene i de forskellige celler meget små. Derfor må man være forsigtig i sine konklusioner. Datagrundlaget må derfor øges betragteligt, før vi kan få fuldt udbytte af alle de tekniske muligheder, der ligger i analyseprogrammet.

Tak

Tak til de to anonyme fagfæller for yderst relevante kommentarer til første udkast af artiklen. Jeg vil også takke redaktørerne af festskriftet for deres tålmodighed og velvilje. Tak til to af mine kolleger på instituttet, Iben Nyholm Debess og Knút Háberg Eysturstein, for teknisk assistance, og Inger Smærup Sørensen skal have tak for sproglige kommentarer.

Litteraturliste

Bugge, Edit. 2018. Attitudes to Variation in Spoken Faroese. I *Journal of Sociolinguistics* 22:3, 1−19. https://doi.org/10.1111/josl.12283.

Bugge, Edit og Jógvan í Lon Jacobsen. 2018. Hugburður til variatión í føroyskum talumáli. I *Íslenskt mál og almenn málfræði*, 40. årg., 97–117. Íslenska málfræðifélagið, Reykjavík.

Chambers, Jack K. og Peter Trudgill. 1990. *Dialectology*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.

Hagström, Björn. 1977. Hví hevur nekarin fepur? I *Fróðskaparrit*, 25. bók, 26–56. Mentunargrunnur Føroya Løgtings, Tórshavn.

Hudson, R.A. 1999. *Sociolinguistics*. Second edition. Cambridge University Press. Cambridge

Navne på stater. Nationalitetsbetegnelser. Dansk-færøsk-islandsk. 1974. *Sprog i Norden*, 81–113. Årsskrift for de nordiske sprognævn. Gyldendalske boghandel. Nordisk forlag, Keypmannahavn.

Petersen, Hjalmar P. 2019. *Føroysk mállæra 1. Kyn, orðmyndan og bending*. Nám, Tórshavn.

Poulsen, Jóhan Hendrik Winther. 2004. *Mál í mæti. Greinasavn eftir Jóhan Hendrik W. Poulsen*. Útgivið í sambandi við sjeyti ára føðingardag hansara 20. juni 2004, redigeret af Anfinnur Johansen og Hans Joensen. Føroya Fróðskaparfelag, Tórshavn.

Sandøy, Helge. 2017. Kjelder og årsaker til dialektendringar i norsk. I *Bók Jógvan. Heiðursrit til Jógvan í Lon Jacobsen á 60 ára degnum*, redigeret af Zakaris Svabo Hansen, Anfinnur Johansen, Hjalmar P. Petersen og Lena Reinert, 371-384. Fróðskapur. Faroe University Press, Tórshavn.

Sandøy, Helge. 2008. Kontakt og spreiing. I *Språkøte. Innføring i sosiolingvistikk*, redigeret af Brit Mæhlum, Gunnstein Akselberg, Unn Røyneland og Helge Sandøy, 221-240. Cappelen akademisk forlag.

Sandøy, Helge. 1994. Stavilsisberandi N í føroyskum. *Málting* 11 (nr. 2, 4. årg.), 12–19.

Weyhe, Eivind. 2015. Færøsk gennem to hundrede år. I *Talemål etter 1800,* redigeret af Helge Sandøy, 409–431. Novus, Oslo.

Weyhe, Eivind. 2012a. Pluralis af nomina agentis på *–ari* i færøsk. I *Eivindaródn. Greinar 1979-2011*, redigeret af Anfinnur Johansen, Turið Sigurðardóttir og Tóta Árnadóttir, 355–365. Fróðskapur, Faroe University Press, Tórshavn.

Weyhe, Eivind. 2012b. Nøkur orð um bendingarmun í føroyskum bygdamálum. I *Eivindaródn. Greinar 1979–2011*, redigeret af Anfinnur Johansen, Turið Sigurðardóttir og Tóta Árnadóttir, 366–373. Fróðskapur. Faroe University Press, Tórshavn.

Weyhe, Eivind. 2012c. Bendingarmunur í føroyskum málførum. I *Eivindaródn. Greinar 1979-2011*, redigeret af Anfinnur Johansen, Turið Sigurðardóttir og Tóta Árnadóttir, 374-419. Fróðskapur. Faroe University Press, Tórshavn.

Internetadresser:

https://ordnet.dk/ods
https://ordnet.dk/ddo
islex.hi.is
malrad.fo/orðalistar/lond og tjóðir

Temporal relations in North Sámi ECM constructions

Marit Julien
Lund university

Abstract

The embedded verb in North Sámi ECM-constructions can appear in one of three different forms: past participle, progressive and infinitive. The existing descriptions of North Sámi say that the past participle places the embedded event before the higher event, that the progressive (traditionally called *aktio essive*) expresses temporal coincidence with the higher event, and that the infinitive normally gets a future interpretation, but it might also coincide temporally with the higher clause. This paper shows that although these generalisations are mostly correct, variation in the temporal interpretation of ECM complement clauses can be caused by a number of factors. In particular, the semantics of the matrix verb and the aspectual properties of the lower verb can influence the temporal relation between the matrix event and the embedded event. In addition, temporal adverbials can shift or fix the time of the embedded event.

Keywords: North Sámi, non-finite clause, tense

1. Introduction[1]

North Sámi has several non-finite verb forms, as can be seen from grammars such as Nielsen (1979), Nickel (1990), Nickel & Sammallahti (2011), Svonni (2015, 2018), and also from Ylikoski (2009). Some of the non-finite forms mainly or exclusively have adverbial uses. The so-called gerund, shown in (1), is an example of this.[2]

(1) Ruoktot mana-dettiin mii garvit dán báikki.
homewards go-GER we avoid.PRES.1PL this.ACC place.SG.ACC
'Going home we avoid this place.'

The temporal interpretation of the gerund is fixed; it denotes an event which coincides with the matrix event. Fixed temporal interpretation is typical of North Sámi non-finite adverbial clauses in general. The temporal properties of non-finite complement clauses are also traditionally considered to be relatively stable. In this paper, I will take a closer look at the temporal interpretations of non-finite complement clauses of one specific type, namely, so-called ECM constructions, i.e. constructions where the matrix verb takes as its only complement a clause with an accusative subject and a non-finite verb, where the latter is either an infinitive, a past participle, or a progressive form.[3,4] After a brief introduction of North Sámi ECM constructions in section 2, I address ECM constructions with embedded past participles in section 3, ECM constructions with embedded progressives in section 4, and ECM constructions with embedded infinitives in section 5. My findings are summarised in section 6. It turns out that although the existing descriptions of

[1] I would like to thank the speakers of North Sámi who shared their judgements with me, as well as two anonymous reviewers who gave very helpful comments to an earlier version of this paper.

[2] The examples in this paper are taken from SIKOR, the Sámi corpus developed by UiT The Arctic University of Norway and the Norwegian Saami Parliament, version 06.11.2018. See gtweb.uit.no/korp/. Some of the examples have been lightly edited, but deviations from the written norm have not been corrected.

[3] ECM is short for "Exceptional Case Marking".

[4] Magga (1986) discusses a variety of structures with embedded infinitives. In the chapter on accusative and infinitive, as he calls it, he also includes constructions that could more correctly be analysed as causatives or as control structures.

 Nordlyd 46.1: 115–124, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6261

the temporal properties of North Sámi non-finite complement clauses mostly holds true of these constructions, all three types allow some variation that also should be recognised.

2. ECM constructions in North Sámi

In North Sámi, the matrix verb in ECM constructions can be a verb of saying and cognition, as in (2), or an experiencer verb, as in (3), including *gávdnat* 'find', shown in (3c).

(2) a. Sii jáhkket álddagas-a cahkkeh-an dola.
3PL.NOM think.PRES.3PL lightning-SG.ACC spark-PAST.PTC fire.SG.ACC
'They think that the lightning sparked the fire.'

b. Son dadjá sin ipmird-it olbmu-id balu.
3SG.NOM say.PRES.3SG 3PL.ACC understand-INF person-PL.GEN fear.SG.ACC
'S/he says that they understand people's fear.'

c. Elle muitala beroštumi lassán-eamen.
Elle tell.PRES.3SG interest.SG.ACC grow-PROG
'Elle says that the interest is growing.'

(3) a. Dovdá earáid šadda-min gierdemeahttum-in.
feel.PRES.3SG other.PL.ACC become-PROG impatient-ESS
'S/he feels that the others are getting impatient.'

b. Báifáhkka gulan muhtim-a boahti-men hoahpu-s.
suddenly hear.PRES.1SG somebody-SG.ACC come-PROG hurry-SG.LOC
'Suddenly I hear somebody coming in a hurry.'

c. Nisu gávnna-i máná-s veallá-min seaŋgga-s.
woman.SG.NOM find-PAST.3SG child.SG.ACC-POSS.3SG lie-PROG bed-SG.LOC
'The woman found her child lying in bed.'

d. Mii leat oaidná-n sin bilid-it luonddu.
we be.PRES.1PL see-PAST.PTC 3PL.ACC destroy-INF nature.SG.ACC
'We have seen them destroy nature.'

In these examples we also see the three forms that the embedded verb in North Sámi ECM constructions can take: the past participle, as in (2a), the infinitive, as in (2b) and (3d), and the progressive (traditionally called *aktio essive*), as in (2c) and in (3a–c).[5] Concerning the distribution of these forms, there is a general consensus in the grammars that the past participle can be embedded under any verb, whereas the infinitive is embedded under matrix verbs of saying and cognition and the progressive is embedded under experiencer verbs (Bergsland 1961:102, Sammallahti 2005:92, Svonni 2018:136, 231).

As for the temporal interpretations of these forms, descriptive grammars and scholarly works on North Sámi all agree that the the past participle places the embedded event before the higher event (Nielsen ([1926] 1979:396, Bergsland 1961:102, Nickel 1990:441). Sammallahti (2005:92), using the term "terminative aspect", states that the past participle indicates that the event denoted by the embedded clause is completed. This is repeated in Nickel & Sammallahti (2011:264). Nickel (1990:440) says that he infinitive normally gets a future interpretation, but it might also coincide temporally with the higher clause. The progressive is said to express temporal coincidence with the higher event (Nielsen ([1926] 1979:384, Nickel 1990:442).[6] Sammallahti (2005:92) and Nickel & Sammallahti (2011:263, 265) state that the infinitive and the progressive both denote non-completed events, so that no distinction is observed between these forms as far as their temporal properties are concerned.

[5] These forms can also appear as main verbs in finite clauses, in combination with various auxiliaries. For reasons of space I will not show examples of this here.

[6] Svonni (2015, 2018) does not comment on the temporal properties of these constructions.

Magga (1986), who discusses the syntax of non-finite clauses with infinitives in much detail, only touches briefly upon the temporal semantics, saying that the past participle correponds to a finite clause in the past or perfect tense, whereas the infinitive and the progressive (called "gerund" in this work) both can refer to the future or coincide temporally with the matrix verb (Magga 1986:175–176).

In the following sections we will see that the generalisations found in the works mentioned above are mostly correct, but also that factors such as the aspectual properties of the lower verb and the semantics of the matrix verb can influence the temporal relation between the matrix event and the embedded event.

3. ECM constructions with past participles

When the verb in a North Sámi ECM complement appears in the past participle form, it denotes an event that precedes the matrix event, according to the grammars. It turns out, though, that the embedded event as a whole does not necessarily precede the matrix event. It depends on whether or not the embedded event is telic. In (4) and (5), where the lower verbs are telic (both are achievements), the embedded event is completed before the time of the matrix event.[7, 8]

(4) Poliissat jáhkket soapmásiid goddá-n geaŧkki.
police.PL.NOM think.PRES.3PL somebody.PL.ACC kill-PAST.PTC wolverine.SG.ACC
'The police think that somebody have killed the wolverine.'

(5) Sara ii loga iež-as bálto-n girdim-is.
Sara.NOM NEG.3SG say.CNG self-3SG.ACC get.scared-PAST.PTC flying-SG.LOC
'Sara says that she has not got scared of flying.'

But in (6), where the embedded verb *bargan* 'worked' is atelic, denoting an activity, it is possible that the activity is still going on at the time of the matrix event. However, the activity must have begun before the matrix event time. Thus, the lower clause gets a universal perfect reading, i.e. a reading where the predicate holds throughout an interval stretching from some time in the past up to the speech time (see e.g. Iatridou, Anagnostopoulou & Izvorski 2003 and the references cited there).

(6) Son dadjá iež-as barga-n turistta-i-guin gosii olles eallima.
s/he say.PRES.3SG self-3SG.ACC work-PAST.PTC tourist-PL-COM almost whole life.SG.ACC
'S/he says that s/he has worked with tourists almost all her/his life.'

In (7), the embedded predicate is stative, and in this case, we are only informed that the state held true at some time in the past, and it is left open whether it still holds:[9]

(7) Son lohká buohkaid diehtá-n dan.
s/he say.PRES.3SG everybody.ACC know-PAST.PTC it.ACC
'S/he says that everybody knew it.'

In the examples (4)–(7) above, the matrix verbs are verbs of speaking and thinking. These verbs do not in themselves place any restrictions on the temporal properties of the embedded predicate. Experiencer verbs

[7] The sentential negation in North Sámi is an auxiliary which has only finite forms. Consequently, non-finite clauses cannot host sentential negation. The negation has to be expressed in the matrix clause instead, as in (5). As this example also shows, the negation combines with the so-called connegative form of the following verb.

[8] The presence of the reflexive pronoun *iežas* in (5) suggests that at least for some speakers, the lower clause must have an overt subject even when it is coreferential with the higher subject. It cannot be phonologically empty, hence not *pro* or PRO. Examples like (i), with an unexpressed embedded subject, can however also be found:

(i) Hansen lohká pláne-me áibbas ođđa hotealla cegge-t.
Hansen.NOM say.PRES.3SG plan-PROG entirely new hotell.SG.ACC put.up-INF
'Hansen says that s/he is planning to put up an entirely new hotel.'

[9] An anonymous reviewer points out that adverbials can have consequences for the temporal interpretations of examples like (7). That is true, but I am interested here in the temporal interpretations arising from the verb forms themselves.

are different in this respect, as we can see in (8) and (9). In (8), where the achievement verb *vuoitit* 'win', in the past participle form, is embedded under the experiencer verb *gullat* 'hear', the resulting reading is that the winning event precedes the hearing event – but what is heard is a report of the winning event, not the event itself.

(8) Son čierru go gullá iež-as vuoitá-n 100 000 ruvnno.
s/he cry.PRES.3SG when hear.PRES.3SG self-3SG.ACC win-PAST.PTC 100 000 crown.SG.GEN
'S/he cries when s/he hears that s/he has won 100 000 crowns.'

Similarly, in (9) the matrix verb is the experiencer verb *oaidnit* 'see', while the embedded verb is the achievement verb *náđustit* 'huddle down', in the past participle form. The interpretation is that the event of huddling down occurred before the seeing event, and consequently, what is seen is the result of the embedded event, not the event itself.

(9) Son oinni-i ealgga náđust-an gieddá-i.
s/he see-PAST.3SG moose.SG.ACC huddle.down-PAST.PTC meadow-SG.ILL
'S/he saw that the moose had huddled down in the meadow.'

More generally, when a non-finite clause with a past participle is embedded under an experiencer verb, the embedded event precedes the matrix event, and the embedded event is not directly experienced.

We can conclude that ECM clauses with past participles denote events that at least partly precede the matrix event. The event that the participle represents can however bear any temporal relation to another embedded event. In (10), the event denoted by the past participle follows the event referred to in the (finite) adverbial clause.

(10) Larsson dadjá iež-as hirpmástuvva-n go Obama oačču-i
Larsson say.PRES.3SG self-3SG.ACC surprise-PAST.PTC when Obama get.PAST.3SG
ráfi-bálkkašumi.
peace-prize.SG.ACC
'Larsson says that he was surprised when Obama got the Peace Prize.'

In (11), on the other hand, the event denoted by the embedded participle precedes the event in the adverbial clause (which in its turn precedes the matrix event).

(11) Su áhčči muitala nieidda-s lávlu-goahttá-n
3SG.GEN father.SG.NOM tell.PRES.3SG daughter.SG.ACC-POSS.3SG sing-begin-PAST.PTC
ovdalgo máhtti-gođii hálla-t.
before know-begin.PAST.3SG speak-INF
'Her father says that his daughter began to sing before she began to be able to speak.'

Finally, in (12) the event denoted by the embedded participle overlaps with the event in the adverbial clause.

(12) Filbma-dahki muitala filmma ráhkad-an go lei
film-maker.SG.NOM tell.PRES.3SG film.SG.ACC make-PAST.PTC when be.PAST.3SG
filbma-skuvllas.
film-school-SG.LOC
'The filmmaker says that (s/he) made the film when (s/he) was at film school.'

In (13) the highest verb, a verb of saying, is in the present tense, and it embeds the past participle of the experiencer verb *gullat* 'hear' which in its turn embeds the past participle of the activity verb *vávjit* 'criticize'. Now the hearing event can be simultaneous with the criticizing event, but both must precede the highest event, the event of telling.

(13) Muitala iežas gulla-n olbmu-id vávjá-n dan.
tell.PRES.3SG self-3SG.ACC hear-PAST.PTC person-PL.ACC criticize-PAST.PTC it.ACC
'S/he says that s/he has heard people criticize it.'

Summing up, the event denoted by a past participle in a North Sámi ECM construction can have any temporal relation to other embedded events, but it must at least partly precede the highest event. If the embedded event is telic, it precedes the higher event in its entirety, but if the embedded event is atelic, the requirement is only that it must have begun before the higher event.

4. ECM constructions with progressive verbs

The embedded verb in a North Sámi ECM construction can however also appear in the progressive form. There is then normally temporal overlap between the event expressed in the higher clause and the event expressed in the complement clause.

As noted in many earlier works on North Sámi (see section 2), experiencer verbs often appear with non-finite complements where the verb is in the progressive form, as exemplified in (14) and (15). In (14), the lower verb denotes an activity, whereas in (15) it is an achievement verb, i.e. a verb denoting a punctual event. In both cases, the event time of the matrix clause is included in the event time of the embedded clause. For (15) this means that the interpretation of the embedded progressive is that the event is durational – stretched out in time. The verb is coerced into this reading by the progressive form.

(14) Gulan olbmuid šurra-me lášmmohallan-lanjas..
hear.PRES.1SG person.PL.ACC chatter-PROG gymnastics-room.SG.LOC
'I hear people chattering in the gymnastics room.'

(15) Jovni ja Liisá oaidniba táksi jávka-me.
Jovni and Liisá see.PRES.3DU taxi.SG.ACC disappear-PROG
'Jovni and Liisá see the taxi disappearing.'

Verbs of saying and cognition can however also embed progressive verbs.[10] There is then also normally temporal overlap between the matrix event and the embedded event. This is illustrated by examples (16) and (17), where the embedded predicates are stative, and the states denoted by the lower predicates must hold at the time of the reported utterances.

(16) Lohká jiena lea-men erenoamáš čielggas-in Operadálu-s.
say.PRES.3SG sound.SG.ACC be-PROG especially clear-ESS opera-building-SG.LOC
'S/he says that the sound is especially clear in the Opera building.'

(17) Sara dadjá iež-as illud-eame gulla-t logaldallam-iid.
Sara.NOM say.PRES.3SG self-3SG.ACC look.forward-PROG hear-INF lecture-PL.ACC
'Sara says that she is looking forward to hearing the lectures.'

When the embedded verb denotes an activity or process, the interpretation is that the activity or process is ongoing at the time of the matrix event. This holds when the matrix verb is in the present tense, as in (18) and (19), and when it is in the past tense, as in (20).

(18) Sylvi lohká iež-as studere-min.
Sylvi.NOM say.PRES.3SG self-3SG.ACC study-PROG
'Sylvi says that she is studying.'

[10] Some speakers prefer an infinitival *leat* 'be' accompanying the progressive in these cases. It has been suggested to me that at least in progressive ECM complements following verbs of saying and cognition, there is always an infinitival auxiliary *leat* 'be' in front of the progressive verb, whether or not it is spelled out. It is striking, though, that in the SIKOR corpus (see fn. 2) there are altogether three occurrences of *leat* between the accusative and the progressive in an ECM complement of the verbs *dadjat* 'say' or *lohkat* 'say'. By comparison, there are 104 cases without *leat*. In order to maintain the idea of an elided *leat*, it must be explained why ellipsis is so dominant in the texts. Sammallahti (2005:150) notes, though, that an infinitival *leat* 'be' is now also sometimes seen with embedded past participles.

(19) Ellen muital-a beroštumi lassán-eamen.
Ellen tell-PRES.3SG interest.SG.ACC increase-PROG
'Elle says that the interest is increasing.'

(20) Eadni várra doaivvu-i mu manna-me hivsseg-ii.
mother.SG.NOM maybe think-PAST.3SG 1SG.ACC go-PROG lavatory-SG.ILL
'Maybe mother thought that I was going to the lavatory.'

In (21), the embedded progressive is an achievement verb, and it gets a habitual reading, again overlapping with the matrix event:

(21) Deanu Hotealla ii loga iež-as vuovdi-min koarttaid..
Deanu Hotel NEG.3SG say.CNG self-3SG.ACC sell-PROG card-PL.ACC
'Deanu Hotel says that they do not sell cards.'

An iterated, generic or habitual reading is of course also possible with activity verbs, as in (22):

(22) Mun loavttán bures go oainnán su dánse-me.
I pass.time.PRES.1SG well when see.PRES.1SG 3SG.ACC dance-PROG
'I have a nice time when I watch him/her dancing.'

It turns out, however, that the time of the embedded event can be shifted forwards or backwards by adverbials, at least when the matrix verb is a verb of saying or cognition. In (23), the matrix verb is in the present tense, but the temporal clause inside the non-finite complement clause denotes an event that is in the past relative to the matrix event, and the event denoted by the progressive *oađđimin* 'sleeping' is taken to coincide with this past event, not with the matrix event.

(23) Son lohká iež-as oađđi-min go telefuvdna riŋgi-i.
s/he say.PRES.3SG self-3SG.ACC sleep-PROG when phone.SG.NOM ring-PAST.3SG
'S/he says that s/he was sleeping when the phone rang.'

Similarly, in (24) the complement clause contains the adverbial *gaskavahkku* 'on Wednesday', and as a result, the progressive *manname* 'going' denotes an event that will take place on Wednesday. In this case, the embedded event is in the future relative to the matrix event.

(24) Nutti lohká iež-as manna-me Álttesjávrá-i gaskavahkku.
Nutti say.PRES.3SG self-3SG.ACC go-PROG Álttesjávri-ILL Wednesday.GEN
'Nutti says that he is going to Álttesjávri on Wednesday.'

Thus, the generalisation that a progressive complement clause denotes an event that coincides with the matrix event is not exceptionless. It seems to hold true of progressive complement clauses that are embedded under experiencer verbs. But when a progressive complement clause is embedded under a verb of saying or cognition, then an adverbial in the lower clause may introduce a time which is different from the time of the matrix event, with the result that the time of the event denoted by the progressive coincides with the time of the adverbial instead. In the absence of elements that shift the time of the lower clause the reading is however one of temporal coincidence.

5. ECM constructions with infinitives

We will now turn to North Sámi ECM constructions with embedded infinitives. These are known to have variable temporal interpretations – the infinitive can denote an event in the future relative to the matrix event, but it can also overlap temporally with that event. The question is then if the choice between the future and the overlapping reading is free, or if it depends on other factors. This is what I will look at more closely in the following.

Contrary to claims in earlier works on North Sámi (see section 2), infinitival ECM complement clauses can be embedded under experiencer verbs. The infinitive will then be interpreted as denoting an

event that overlaps with the matrix event. This holds regardless of whether the lower verb is stative, as in (25), punctual, as in (26), or dynamic and durational, as in (27).

(25) Son dovdá iež-as máhtti-t sáme-giela bures.
s/he feel.PRES.3SG self-3SG.ACC know-INF sámi-language.ACC well
'S/he feels that s/he knows the Sámi language well.'

(26) Mii oinniimet Norgga vuoiti-t Brasiil-a badjel.
we see.PAST.1PL Norway.ACC win-INF Brazil-GEN over
'We saw Norway beat Brazil.'

(27) Lei máilmmi somá beassa-t gulla-t Elina juoiga-t.
was very fun get.to-INF hear-INF Elin.ACC juoigat-INF
'It was great fun getting to hear Elin *juoigat*[11].'

The example in (28) appears to be an exception, since the embedded clause denotes a future event relative to the matrix event, although the matrix verb is *oaidnit* 'see'. However, *oaidnit* is used as a cognition verb here, just like *see* in the English translation. Thus, (28) is no exception after all, since infinitives embedded under verbs of cognition can have a future interpretation relative to the matrix verb. In this particular case the future interpretation stems from the verb *šaddat*, which often is a futural verb.

(28) Muhto oainn-án dan šadda-t váttis-in.
but see-PRES.1SG it.ACC become-INF difficult-ESS
'But I see that it will be difficult.'

The availability of a future reading of an infinitival clause embedded under a verb of saying or cognition depends partly on the lexical aspect of the embedded verb. When an embedded infinitival verb denotes a state, this state overlaps temporally with the higher event. This is illustrated in the examples below. The verbs *guoskat* 'concern' in (29), *dovdat* 'feel' in (30), *liikot* 'like' in (31) and *dárbbašit* 'need' in (32) are all stative, and in each case, the state overlaps with the matrix event. More precisely, the time of the matrix event is included in the time span in which the state holds.

(29) ON-komitéa dadjá dán guoska-t eami-álbmog-i-idda.
UN-committee.SG.NOM say.PRES.3SG this.ACC concern-INF indigenous-people-PL-ILL
'The UN committee says that this concerns indigenous peoples.'

(30) Son maiddái dadjá iež-as dovda-t beahtahallan.
s/he also say.PRES.3SG self-3SG.ACC feel-INF disappointed
'S/he also says that s/he feels disappointed.'

(31) Sii dadjet iež-aset liiko-t sudno musihkki-i.
they say.PRES.3PL self-3PL.ACC like-INF 3DU.GEN music-SG.ILL
'They say that they like their music.'

(32) Son lea čielgasit dadja-n iež-as dárbbaš-it bargo-ráfi.
s/he be.PRES.3SG clearly say-PAST.PTC self-3SG.ACC need-INF work-peace.SG.ACC
'S/he has said clearly that s/he needs peace to work.'

When an embedded infinitival verb has a habitual or generic reading, there is also temporal overlap with the higher event. This is not surprising, since we know that habituals and generics are stative (Krifka et al. 1995). In (33)–(37) below, the embedded clauses all have infinitival verbs. The embedded clauses in (33)–(36) represent habitual events, whereas the embedded clause in (37) is generic. They all denote events that coincide temporally with the matrix event.

[11] To *juoigat* is to sing in the traditional Sámi style.

(33) Nils lohká iežas hárjehalla-t spáppa čiekča-t
Nils.NOM say.PRES.3SG self-3SG.ACC practice-INF ball.SG.ACC kick-INF
njeallje geardde vahku-s.
four.SG.GEN time.SG.GEN week-SG.LOC
'Nils says that he practices football four times a week.'

(34) Ánde dadjá iež-as vuoiŋŋast-it go beassá juoigga-st-it.
Ánde.NOM say.PRES.3SG self-3SG.ACC relax-INF when get.to.PRES.3SG juoigat-DIM-INF
'Ánde says that he relaxes when he gets to *juoigat* a little.'

(35) Vuoddji loga-i iežas barga-t fabrihka-s.
driver.SG.NOM say-PAST.3SG self-3SG.ACC work-INF factory-SG.LOC
'The driver said that s/he worked in a/the factory.'

(36) Ollugat dadjet iež-aset lohka-t ja čálli-t sáme-giela
many.PL.NOM say.PRES.3PL self-3PL.ACC read-INF and write-INF Sámi-language-SG.ACC
hui bures.
very well
'Many say that they read and write Sámi very well.'

(37) Olstad ii oainne vearjju-id čoavdi-t maidege.
Olstad.NOM NEG.3SG see.CNG weapon-PL.ACC solve-INF anything.ACC
'Olstad doesn't think that weapons solve anything.'

But apart from habitual and generic uses, the interpretation of dynamic infinitival verbs in complement clauses embedded under verbs of saying and cognition depends on the context. In (38), the adverbial *boahtte jagi* 'next year' makes it clear that a future interpretation of the embedded verb *divrut* 'become (more) expensive' is intended.

(38) Čeahpi-t doivot bensiinna divr-u-t vel eambbo
expert-PL.NOM think.PRES.3PL petrol.SG.ACC expensive-INCH-INF even more
boahtte jagi.
coming year.SG.GEN
'Experts think that petrol will get even more expensive next year.'

In the absence of temporal adverbials, however, the interpretation of embedded infinitival clauses mostly depends on world knowledge in combination with the overall lexical context. In (39), the most plausible interpretation is one where the event denoted by the embedded infinitival clause coincides with the matrix event, whereas in (40), on the preferred interpretation the embedded event is in the future relative to the matrix event:

(39) In jáhke du váldi-t mu sávaldaga-id bealljái-ge!
NEG.1SG think.CNG you.ACC take-INF my wish-PL.ACC ear.SG.ILL-even
'I don't think that you are even listening to my wishes.'

(40) Eamit ballá isid-a jápmi-t giddagas-as.
wife.SG.NOM fear.PRES.3SG husband-SG.ACC die-INF prison-SG.LOC
'The wife fears that the husband will die in prison.'

In (41), the matrix clause is in the past perfect, and the embedded infinitival clause denotes an event which is in the future relative to the past matrix event, but again, this is more due to the lexical content than to the grammar:

(41) In lea-n jáhkká-n Kristina váldi-t medálja.
NEG.1SG be-PTC think-PTC Kristina.ACC take-INF medal.SG.ACC
'I had not thought that Kristina would take a medal.'

In (42), the embedded clause is ambiguous between a simultaneous reading and a futural reading, and it turns out that the preferences vary between speakers, based on verb meanings and possibly also on the demonstrative:

(42) Mun doaivvun min lihkostuvva-t dainna.
I hope/think.PRES.1SG 1PL.ACC succeed-INF it.COM
'I hope/think that we will succeed/are succeeding with it.'

When it comes to the example in (43), however, there are no preferences. The embedded infinitival clause is completely ambigous – it can refer to a present situation or to a future situation – and there are no lexical or other clues that could lead speakers to favour one interpretation over the other. Only more context could disambiguate the clause.

(43) Ballet mirkku-id golga-t johki-i.
fear.PRES.3PL poison-PL.ACC flow-INF river-SG.ILL
'They fear that the poisons are flowing/will flow into the river.'

Summing up, we have seen that in ECM constructions with an infinitival verb embedded under an experiencer verb, the interpretation is necessarily one of temporal coincidence. If the matrix verb is a verb of saying or cognition and the embedded infinitival verb is stative, there is also temporal overlap between them. But if the matrix verb is a verb of saying or cognition and the embedded infinitival verb is dynamic, then the infinitival verb can denote an event that overlaps with the matrix event or an event that is in the future relative to the matrix event. The infinitival verb in itself is ambiguous, and in cases where one temporal interpretation is favoured over the other, this follows from the linguistic and/or non-linguistic context.

6. Conclusions

In North Sámi, ECM complements can be selected by verbs of saying and cognition and by experiencer verbs. Three verb forms can appear in these complements: the past participle, the progressive and the infinitive. In this paper I have looked more closely at the temporal interpretations that these forms give rise to and compared my findings to the descriptions found in the grammars of the language.

Concerning the past participle in ECM constructions, the grammars say that it denotes an event that precedes the matrix event. It turns out, though, that this is strictly correct only in cases where the past participle is a telic verb. If the past participle instead is an atelic verb, the event that it denotes may still be ongoing at the time of the matrix event, although it must have begun before the matrix event.

ECM complements with progressive verbs are said to denote events that coincide temporally with the matrix event. This holds true of progressive verbs embedded under experiencer verbs, but if the matrix verb is a verb of saying or cognition, the time of the embedded event can be shifted forwards or backwards by adverbials, giving an interpretation where the embedded progressive is not simultaneous with the matrix event. The embedded event overlaps with the matrix event only in the absence of such adverbials.

The temporal interpretation of infinitival complement clauses depends on the semantic class of the matrix verb and on the lexical aspect of the embedded verb. If the matrix verb is an experiencer verb, there is temporal overlap between the matrix event and the embedded event. This is also the reading that arises if the lower verb is stative, or if it is habitual or generic, irrespective of the properties of the matrix verb. If the matrix verb is a verb of saying or cognition, there is temporal overlap with embedded stative verbs, and with verbs with habitual or generic interpretation. But if a dynamic infinitival verb is embedded under a verb of saying or cognition, then the embedded verb can have a simultaneous or a future reading relative to the matrix verb. Only the context, linguistic or non-linguistic, can serve to select one reading in such cases.

References

Bergsland, Knut. 1961. *Samisk grammatikk: med øvelsesstykker*. Kirke- og undervisningsdepartementet, Oslo.

Iatridou, Sabine, Anagnostopoulou, Elena & Izvorski, Roumyana. 2003. Observations about the form and meaning of the Perfect. In *Perfect Explorations*, edited by Artemis Alexiadou, Monika Rathert & Arnim von Stechow, pp. 153–204. De Gruyter Mouton, Berlin. https://doi.org/10.1515/9783110902358.fm

Krifka, Manfred, Pelletier, Francis Jeffrey, Carlson, Gregory, ter Meulen, Alice, Link, Godehard & Chierchia, Gennaro. 1995. Genericity: An introduction. In *The Generic Book*, edited by Gregory N. Carlson & Francis Jeffrey Pelletier, 1–124. Chicago University Press, Chicago.

Magga, Ole Henrik. 1986. *Studier i samisk infinitivsyntaks*. Sámi Instituhtta, Guovdageaidnu.

Nickel, Klaus Peter. 1990. *Samisk grammatikk*. Universitetsforlaget, Oslo.

Nickel, Klaus Peter & Pekka Sammallahti. 2011. *Nordsamisk grammatikk*. Davvi Girji, Kárášjohka.

Nielsen, Konrad. [1926] 1979. Lærebok i lappisk (samisk): utarbeidet på grunnlag av dialektene i Polmak, Karasjok og Kautokeino: grammatikk, tekster og glossar. Vol. 1: Grammatikk: lydlære, formlære, orddannelseslære og syntaks samt tillegg. Brøgger, Oslo. Reprint Universitetsforlaget, Oslo.

Sammallahti, Pekka. 2005. *Láidehus sámegiela cealkkaoahpa dutkamii.* Davvi Girji, Kárášjohka.

Svonni, Mikael. 2015. *Davvisámegiela: sánit ja cealkagat*. Ravda Lágadus, Giron/Kiruna.

Svonni, Mikael. 2018. *Modern nordsamisk grammatik*. Ravda Lágadus, Giron/Kiruna.

Ylikoski, Jussi. 2009. *Non-finites in North Saami*. Suomalais-Ugrilainen Seura, Helsinki.

You can't suggest that?! Comparisons and improvements of speller error models

Heiki-Jaan Kaalep, Flammie Pirinen, Sjur Nørstebø Moshagen
Tartu ülikool (Kaalep), UiT Norgga árktalaš universitehta (Pirinen, Moshagen)

Abstract

In this article, we study correction of spelling errors, specifically on how the spelling errors are made and how can we model them computationally in order to fix them. The article describes two different approaches to generating spelling correction suggestions for three Uralic languages: Estonian, North Sámi and South Sámi. The first approach of modelling spelling errors is rule-based, where experts write rules that describe the kind of errors that are made, and these are compiled into a finite-state automaton that models the errors. The second is data driven, where we show a machine learning algorithm a list of errors that humans have made, and it creates a neural network that can model the errors. Both approaches require collections of misspelling lists and understanding its contents; therefore, we also describe the actual errors we have seen in detail. We find that while both approaches create error correction systems, with current resources the expert-built systems are still more reliable.

Keywords: Spell-Checking, rule-based, fsa, machine learning, sámi languages, estonian

1. Introduction

The ultimate speller only accepts correct words, finds all spelling errors, and always gives the one and only relevant suggestion. This speller will never exist, but it is the ultimate speller we strive to achieve. In this article we explore a few ideas in that direction, and apply them to three languages found in the *GiellaLT* infrastructure[1]: North Sámi, South Sámi and Estonian. More precisely, this article looks at the error model, and how to improve the suggestions given.

To that end, our goal is to reduce the noise level (increase precision) by generating as few irrelevant suggestions as possible, and when in doubt, give no suggestion at all rather than risk giving irrelevant suggestions; this is in contrast with e.g. Hunspell[2] (Trón et al. (2005)) and the rest of the Xspell family (Ispell, Aspell[3], Myspell, nuspell[4], etc). While pursuing this goal, we try to understand the reasons behind mistyping, and assume that classifying the errors will give us some insight. Having this insight, it might be possible to find ways for increasing recall as well.

An attempt to find regularities in misspellings naturally invokes the idea that one might try machine learning for this purpose; one should use all tools available for achieving one's goal.

The approaches that will be investigated are the following:

- hand-crafted regex error model
- machine-learned error model

The work described in this article says nothing about coverage, i.e. how many words flagged by the speller are real errors and how many are actually correct words, missing from the speller's vocabulary; or how many misspelled words are falsely recognized as correct. We limit ourselves to real misspellings.

The article is organized as follows: first, there is a short overview of earlier work. Following that, we'll describe the methods used for developing new error models. We then describe the misspelling lists used for development, testing and evaluation. After that we say a few words about the types of errors in these lists, followed by a short description of the main features of the languages and their orthography, focusing on the parts relevant to this paper. We then describe the new error models in detail, starting with a short overview of our baseline error model, after which we evaluate the performance of the new error models. Finally, there is a discussion on the outcome, and a conclusion.

[1] https://giellalt.github.io/
[2] https://hunspell.github.io/
[3] http://aspell.net
[4] https://nuspell.github.io

 Nordlyd 46.1: 125-139, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6349

2. Earlier work

A lot of work has been done on spelling corrections—we give an overview of the literature here—although most of it looks at English and closely or typologically related languages. See e.g. Kukich (1992), Hládek et al. (2020). Working with languages with a complex morphology and phonology does offer some additional challenges, and minority and indigenous languages with a recent writing culture adds to that challenge, also, not a lot of work has been done in this area.

Finite-state language models have been used in spell-checking and correction for a while, one of the most recent approaches that is the basis of our system as well is Pirinen et al. (2014). Within the Sámi language context, the work has been done from Gaup et al. (2005) onwards.

Substantial work on analysing North Sámi spelling errors was done in Antonsen (2013), and the insights gained were important for the work done with the North Sámi speller in this article. To the best of our knowledge, no other Sámi languages have been analysed with regard to spelling errors, their classification and frequency.

Estonian spelling errors, that emerge while typing on a computer keyboard, have not been described in publications. However, the Estonian spellers that were created by Filosoft Ltd. in the beginning of the 1990ies (e.g. for Microsoft Word) contain a suggestion module, and since their *C*-language source code has been made public[5], it has been possible to re-implement it as an FST.

There is some prior work done on the general problem of error-correction using neural networks and this is often suggested as the state-of-the-art currently, so we have chosen to experiment on this approach as well. In Li et al. (2020) the authors use a neural model to determine the context of the word, resulting in a better guess as to what was the word that the author wanted to use.

One of our central themes in this article lies in the usage and importance of a public error corpus and/or list; an elaborate model for ordering correction candidates: c.f. Flor et al. (2019). Different sources have different types of errors, thus different strategies should be used, and different recall-precision figures are expected: Beeksma et al. (2018).

The GiellaLT framework (Moshagen et al. 2013) originated from the initial work on proofing tools and morphological analysers for the Sámi languages, where Trond Trosterud has been a major driving force (see e.g. Moshagen and Trosterud (2005) and Trosterud and Wiechetek (2007)). The framework itself is language independent, but favours rule-based technologies suitable for morphology rich, complex, and low-resource languages. The overall goal is to support all language technology needs of indigenous and minority languages, from text input to speech technology. It is constantly being developed, and is the home for keyboards for 50 languages, and language models for more than 130 languages. Many languages and keyboards are in daily use, and is core to the digital life of several indigenous and minority language communities.

3. Methods

In this article we study two approaches to error-correction, a rule-based method using two-level *finite-state transducers* (FST) (Pirinen et al. 2014), and data-driven *neural network-based* (NN) (Hochreiter and Schmidhuber 1997, Bollmann and Søgaard 2016) *language models*. We call a method that corrects incorrect word-forms into correct ones *an error model*.

3.1. FST methods

The finite-state spelling correction follows the model described in Pirinen and Hardwick (2012): a transducer that modifies the erroneous string is composed with the speller transducer, which accepts only valid wordforms. As a result, the suggestion transducer presents only modifications that are also valid wordforms to the user.

Ideally, there would be only one suggestion, and this would be the right one. The more suggestions there are, and the lower down the ranked list the correct one is, the worse for the user; and the worst case is a long list of suggestions without the correct one amongst them. So the suggestion transducer has a dual goal: keep the number of the suggestions low, and rank them correctly. One may ask whether it is better to provide no suggestion at all than to present the correct one ranked as 9th, for example. Presently, we have no answer to this question. What are the psychologically comfortable

[5] https://github.com/Filosoft/vabamorf

number and way of ranking, is a question for future research on user studies; presently we just notice that this aspect has to be taken into consideration.

Limiting the number of suggestions can be achieved by either allowing fewer modifications of the erroneous form, limiting the recognizable vocabulary of the speller, or both. As an example: fewer modifications might mean that only edit distance one is allowed, and limited speller vocabulary might mean that only simplex words are allowed, while productively formed compounds are prohibited as suggested corrections[6].

With weighted transducers, we may attach different weights to different edit operations and recognized word-forms. For example, interchanging *d* with *t* adds a certain weight, and every component of a compound word adds another weight. Suggestion ranking will follow from adding up all these weights, and limiting their number may be based on cutting the list either above some absolute weight, or above some absolute number of candidates. However, it is not obvious how one should determine the right final weights and cutting points. This article concentrates on modifications of the erroneous wordforms: what kind of modifications should be made, and whether we can argue for attaching certain weights to these modifications, in order to signal their likelihood.

Weights from the speller lexicon are also used: if two candidates result from modifications with the same weight, then the one which gets smaller weight from the speller is ranked first. We achieve this by having the modification weights surpass the speller weights by a large margin; it is the modification which is important, not the likelihood of the wordform itself. The speller lexicon weights are partly based on frequency of words either in a corpus or by linguistic intuition, and partly on expert-decided likelihood of the morphological tags; more elaborate weighting schemes can be imagined, but that is outside the scope of this article.

3.2. NN methods

For neural error correction modelling, we are using a neural machine translation approach. Within the neural machine translation framework, we use the incorrectly written word-forms as source language, and the corrected word-forms as target language. This logic allows us to train an error correction model with an off-the-shelf neural machine translation toolkit. For this experiment we are using OpenNMT-py[7] (Klein et al. 2017) in its default settings, i.e. a translation model following the OpenNMT tutorial on their website[8].

To limit the creativeness of neural suggestions, we restrict the corrections to word-forms that are acceptable by the dictionary of the rule-based spell-checker. That is, we take the list of n-best translations from OpenNMT-py and check it against the speller lexicon. Only the suggestions accepted by the speller are included in the final suggestion list.

4. Lists of misspellings

It is a truism that texts differ, depending on who creates them, for what purpose and for what readership. Likewise, it is only natural to expect that the errors made while writing depend on various factors. We are aware that the misspelling lists we have at hand are not representative of the "general text class" created by an "average writer"; so, in order to remain cautious when interpreting our results, here are the main characteristics of the corpora that these lists are derived from.

4.1. North Sámi

The present day North Sámi orthography is from 1979, with some smaller adjustments from 1985[9]. The present orthography is thoroughly described in Nickel and Sammallahti (2011).

As a result of the Norwegian assimilation policy towards the Sámi people throughout a major part of the 20th century, it is clear that most texts written in the modern orthography are pretty recent. Modern North Sámi literacy is correspondingly young, which is reflected in texts in the form of spelling and other grammatical errors. In the material

[6] They would still be accepted by the speller. The core idea is that one can use two different transducers or automata for the speller: one to verify the text, including productive morphology, and another, more restricted transducer, to verify suggestions.

[7] https://opennmt.net/OpenNMT-py

[8] https://opennmt.net/OpenNMT-py/quickstart.html

[9] There have been several older orthographies going back all the way to 1748.

used in Antonsen (2013) there is about 4% spelling errors, which is considerably more than in e.g. Norwegian or English texts produced by native speakers. In Flor et al. (2015), where the majority of the texts are written by non-native speakers of English at various levels of mastering the language, the average number of spelling errors is 2.74%. And for the most advanced writers contributing to the data set, the average number of misspellings is well below 1%. That is, the average number of spelling errors in North Sámi texts is considerably higher than in similar English texts. This is expected given the short history of the orthography, the sociolinguistic setting, the paucity of available text and thus written language exposure, and the minority language status of North Sámi.

The material used in developing, testing and evaluating the error models in this paper has been collected over many years while developing various language technology tools for North Sámi.[10] Misspellings found in texts have been collected in a separate text file, together with the expected correction (usually based on the incorrect word form itself, sometimes also considering the context where the misspelling was found). By the time of writing, the list of typos contains 11 706 entries. Since the focus of research described here is evaluating and developing error models, the list was filtered by removing multiword expressions, false negatives[11], and entries for which the given correction was not recognized by the speller. The filtered list consists of 10 745 entries.

Given the development history of the list of typos, the source texts for the misspellings can be assumed to be all sorts of texts, the majority of which are found in SIKOR[12]. That is, the collection of typos can be considered relatively representative of errors made by North Sámi writers of various genres.

For the machine learning experiment, the list was split in three according to the usual 80-10-10: 80% for training, and 10% each for testing and development / validation. For the regular expression experiment, no such split was used, and the list was both used to inform the developers about useful patterns, and to evaluate the resulting error model.

4.2. South Sámi

The present day South Sámi orthography was formally decided upon in 1978, although Bull and Bergsland (1974) used an early version of that orthography. South Sámi differs from most other Sámi languages and dialects due to a vast and complex system of umlaut, c.f. Bergsland (1994) and Magga and Magga (2012). Although South Sámi does not have consonant gradation as opposed to the other Sámi languages, it does have alternations in consonant clusters and surrounding vowels depending on the syllable and foot structure of the word. Various inflectional endings add zero, one or more syllables to the base form, which forces a recast of the foot structure, which can set off a chain reaction of various consonant and vowel changes. Two examples:

(1) a. gåetie *gåetie+N+Sg+Nom* gåatan *gåetie+N+Sg+Ill* gåatetje *gåetie+Dimin+N+Sg+Nom* gåatatjasse *gåetie+Dimin+N+Sg+Ill*
'House, into the house, little house, into the little house'

b. åeruve *åeruve+N+Sg+Nom* åerievasse *åeruve+N+Sg+Ill* åerievadtje *åeruve+Dimin+N+Sg+Nom* åerievadtjese *åeruve+Dimin+N+Sg+Ill*
'Squirrel, into the squirrel, little squirrel, into the little squirrel'

That is, the vowel of the second and third syllables changes as follows: *-ie-, -a-, -e-, -a-* for *gåetie*, and *-u-* + *-e-, -ie-* + *-a-* for *åeruve*. The default illative case ending has two forms: *-asse* and *-ese*, and the diminutive derivation also has two forms: *-etje* and *-adtje*. The form of the suffixes (illative and diminutive in example 1) are solely dependent on the syllable count, whereas some vowel changes also depend on the stem type. The umlaut of the root vowel is triggered by the underlying vowel of both case and derivational suffixes.

The South Sámi language community is just a fraction of the North Sámi, and with correspondingly less production and exposure to the written language. Also, a considerable portion of the population is in practice L2 speakers. This is reflected in the misspelling list used for testing as a number of errors relating to mixing vowel and inflectional endings, essentially miscounting the syllables and thus applying the wrong suffix; an example of this taken from the list can be seen in (2). (2) also contains other errors, like using *ø* for correct *ö*, and mixing *s* and *sj*. Identifying each and every such case reliably is not trivial, identifying the proportion of these errors to the rest is left as a topic for future research.

[10] Source code at: https://github.com/giellalt/lang-sme
[11] misspellings accepted by the speller as valid words.
[12] Giellatekno and Divvun (2021)

(2) a. *Vyøhkesadtibie
vyöhkesjadtedh+V+IV+Ind+Prs+Pl1
'We help each other' (wrong syllabification and thus suffix form)

b. Vyöhkesjadtebe
vyöhkesjadtedh+V+IV+Ind+Prs+Pl1
'We help each other' (correct syllabification and suffix)

Identifying the syllabic structure is not made easier by historic processes leading to exceptions, so that instead of the regular pattern $2 + 2 + \cdots n \cdots + 2/3$, you get $3 + 2$, or $2 + 1$, instead of the expected $2 + 3$, and 3. Examples of these can be seen in (3).

(3) a. dåerie•dieh
dåeriedidh+V+TV+Ind+Prs+Pl3
'They are following' (syllable structure: 2 + 1)

b. dåerede•minie
dåeriedidh+V+TV+Ger
'(In the process of) following' (syllable structure: 3 + 2)

Complicating the issue further are loan words: how should their syllables be counted and fit into the foot structure of South Sámi phonotactics? An example of this can be seen in (4), with the misspelled form in (4a), and the correct form in (4b). It is very clear that the misspelling of the case suffix is caused by applying a wrong foot structure to the word form.

(4) a. Wikipe•dij:ese
wikipedije+N+Sg+Ill
'Into Wikipedia' (wrong syllable structure: 3 + 3, and thus wrong suffix form)

b. Wiki•pedi•jasse
wikipedije+N+Sg+Ill
'Into Wikipedia' (Correct syllable structure: 2 + 2 + 2)

Finally, the South Sámi orthographic rules recommend that one uses Norwegian *æ* and Swedish *ö*. Up until recently, following these rules require that one knows how to produce the vowel letter from the other side of the border, and it also requires an extra key press: AltGr + the standard vowel. In practice, most people didn't care, and the South Sámi list is full of Norwegian *ø*'s and Swedish *ä*'s. These are considered misspellings by the spelling checker, and they also contribute to the complexity of correcting South Sámi. It is not uncommon to find spelling errors with an editing distance of four and more; in the test list of typos 48 such cases are found, ≈4.2% of the corpus.

As was the case with North Sámi, the list of typos for South Sámi is collected while developing the morphological analyser, based on material that is mostly found in SIKOR (Giellatekno and Divvun (2021)). The cleaned version of that manually built list mentioned above contains only 1 154 entries. A separate list of typo-correction pairs was extracted from a manually marked up corpus of gold-standard text. That token list contains 8 325 non-unique entries, and was used for training a machine learning model, testing and evaluation, using the common 80-10-10 split. This list, extracted from the gold standard corpus, was not used when building the manually crafted regex error model.

4.3. Estonian

Estonian orthography in its present form was adopted during the third quarter of the 19th century. It is modelled after Finnish orthography; the proposal was made by Adolf Ivar Arwidsson (1822). Prior to this, Estonian orthography was modelled after High German, but uneducated Estonian peasants spontaneously tended towards the Finnish style orthography (Kask 1970:p. 204)

The main difference from the previous orthography lies in the simplicity of the rules for marking phone length: nowadays, the rule of thumb is that a short phone is marked by one letter, a long (and extra-long) phone by two letters, and every consonant in a cluster is marked with one letter, even if it is pronounced long or extra long. As an exception,

k, *p* and *t* are written as *g*, *b*, *d* when short, *k*, *p*, *t* when long, and *kk*, *pp*, *tt* when extra long. Also, when adjacent to a nonsonorous consonant, *g*, *b*, *d* are also written as *k*, *p*, *t*. In addition to indeterminacy in differentiating between long and extra long phones (except for *k*, *p*, *t*), and between short and long ones in consonant clusters, palatalisation is also not marked. There have been numerous propositions to improve the Estonian orthography, in order to make it even more phonetic, e.g. by allowing double letters in consonant clusters, and three letters for extra long phones, but these propositions have not been adopted. Very succinct hearing and marking of phone lengths is difficult to implement in practice, given the various co-articulation effects in real speech.

In addition to the principle of phone length and letter correspondence, the Estonian orthography also to some extent follows the principle of keeping the traditional form of words (even if it deviates from the current pronunciation), and the principle of retaining the form of morphemes while inflecting the word (Erelt et al. 2007). Orthography errors tend to happen when these two additional principles collide with the phonemic principle.

The Estonian list of 3000 misspelled words originates from journalists' texts. About one third of it dates from the 1980-1990ies: 1) a re-typed-in Corpus of Estonian Literary Language[13], containing 1 million words from 1983–1988, and 2) texts from the news agency Baltic News Service, from one month in 1996 (about 250 000 words). The errors were gathered by running an Estonian morphological analyser on the corpus; and then manually picking misspellings from the set of unanalysed words (by Heili Orav and Leho Paldre). Another two thirds date from 2000-2010ies, gathered by Kairit Sirts from a newspaper corpus in an ad hoc manner, according to her own words.

5. Error types

An ideal error typology would reflect what went wrong in the chain of actions of the writer, and/or what was the likely cause, not just count the edit operations. However, we have not been able to reach this ideal yet. It seems though that one potential distractor might be the current set of conventions for writing the language, i.e. its orthography.

The full list of registered typos was run through a semi-automatic classification system, and tagged according to identified class. The resulting classification combines edit distance with character classes that are involved and is summarized in Table 1. In cases where subclasses are identified, the figures for those are listed to the left in each column, the total to the right.

Accented letter errors are easy to correct: there are very few alternatives one should offer, and the reasoning behind the suggestions is transparent, making it easy for the writer to decide whether to accept or not. An example for Estonian would be **tshempion - tšempion*. For North Sámi, this type of errors is very frequent - one third of misspellings belong to this class, and we can even identify subclasses: vowel *á* vs *a* (e.g. **Amerihka - Amerihká*), or consonants *č*, *đ*, *ŋ*, *š*, *ŧ*, *ž* vs *c*, *d*, *n*, *s*, *t*, *z* (e.g. **Cuovvovaccat - Čuovvovaččat*, **Sámediggerádi - Sámediggeráđi*, **CD-singel - CD-siŋgel*, **oktašas - oktasaš*, **olbmoŧ - olbmot*, **gazaldaga - gažaldaga*). In fact, *a-á* confusion is the single most frequent spelling error in North Sámi texts, around 40% in general according to (Antonsen 2013:p 24)[14]. The source of these errors in North Sámi is likely several. One is lack of keyboard support that makes it hard to type the correct letter. That was a major issue in social media texts investigated by Antonsen op.cit., but for several years now there has been available a North Sámi keyboard app for mobile phones, so this is less of a problem today. Another possible source is insecurity in the correct spelling, often in combination with dialectal variation. The *a-á* confusion can at least partly be attributed to the fact that the orthography does not follow the phonology in various dialects, the variation is greater and more complex than the orthography reflects. Also final *ŧ* instead of final *t* is most likely based on pronunciation: in some dialects, the plosive *t* is reduced to a pure fricative *h* sound when followed by a word beginning with a vowel. As almost all misspellings of *ŧ* for correct *t* can be found in this position, it is very likely that phonology plays a role. For a more detailed analysis of spelling errors in North Sámi, see (Antonsen 2013).

Accented letters in South Sámi covers only three pairs: *i* vs *ï* (e.g. **jih - jïh*, **hïjven - hijven*), **ø* vs *ö* (e.g. **bøøremes - böörems*), and **ä* vs *æ* (e.g. **nännoste - nænnoste*). But they cover more than half of all misspellings in our test data. Out of a total set of 8 325 misspelling instances, 4 285 – or 51.5% – are errors of this type. The conjuction *jïh* (=*and*) alone counts for more than 10% (884 occurrences) of all misspellings. The three pairs fall into two categories, one purely orthographic, and one phonological. The **ø/ö* and **ä/æ* pairs are purely orthographic: as South Sámi is spoken in both Sweden and Norway, the idea is to make a compromise such that one sound is written using a Swedish letter (*ö*) and one using a Norwegian letter *æ*. Due to the lack of a South Sámi keyboard, people have usually fallen back to using either a Norwegian or a Swedish keyboard, disregarding the orthographic norm. In the case

[13] https://www.cl.ut.ee

[14] she includes real-word errors, which we do not, which probably explains the difference in relative size for this error type in her investigation compared to our findings.

Main error class	Subclass	Estonian	North S	South S
	á vs *a*		25	
	čđŋšžŧ vs *cdnstz*		8	
Only accented letter errors		2	= 33	5
	double or diphthong	7	13	
	other	37	7	
Delete 1		= 44	= 20	18
	double or diphthong	7	11	
	other	16	4	
Add 1		= 23	= 15	14
Substitute 1		13	17	11
	1 to 2 or 2 to 1		3	7
	adjacent		2	4
Substitute 2		0	= 5	= 11
Transposition		10	2	2
Repetition; South S=suffix		2	0	3
Other		6	8	36
Total		100%	100%	100%

Table 1: Error types, percentage of all errors.

of *i* vs *ï* it is a real phonological opposition, although the distinction was not made in early versions of the South Sámi orthography. The distinction is also not clear to all speakers.

As seen above, the error type **accented letters** is a heterogenous class, with various properties across the languages. It still makes sense to treat them as one with respect to modelling errors, as they stand out from other misspellings both in frequency and often simplicity of correction.

Deleting (or omitting) a letter is a very frequent error. It may be caused by failing to hit a key, or by failing a phone-to-letter mapping rule. A suggestion to correct this error by doubling a letter, or (in case of North Sámi) by creating a diphthong, e.g. **departementa - departemeanta*, might seem more plausible than a suggestion to insert a letter in some random position of the same word. Thus, it makes sense to identify this subclass of deletions.

If the misspelling means that an extra letter has been **added**, we also identify a subclass of resulting doubles or diphthongs, the classification thus being similar to the deletion errors.

Substitution errors are relatively more frequent in the Sámi corpora than in Estonian. They also involve cases where one letter is substituted by two (e.g. North Sámi **direktora - direktevra*), or two by one (e.g. North Sámi **Osllu - Oslo*), or two adjacent letters by two different ones, as in consonant gradation mix-ups (e.g. **Sámedikkeválgii - Sámediggeválgii*).

In Estonian, the main source of errors is the typing process, as evidenced by the relatively high proportion of **transpositions** (e.g. **komapnii - kompanii*) and repetitions (e.g. **poliititika - poliitika*). Errors relating to incorrectly writing phones are relatively few. In North Sámi, the main source of errors is the phone-to-letter process, i.e. applying rules of orthography. Many substitution errors may be blamed on it. This is also documented and discussed by (Antonsen 2013).

In South Sámi as well, the main source of errors is the phone-to-letter process, i.e. applying rules of orthography. In addition, another major source of error is the morphophonology of the language, especially as related to syllable structure and its consequences for **suffix** realisation, as exemplified by **edtjibie* vs *edtjebe*. But the biggest class of errors in South Sámi is the unclassified **other** group — these are typos that are not easily classified by the means used in this work.

6. Error models

The error models we study are: the baseline, a new regex model, and a machine learned model. The baseline model is a general edit distance 2 model built from the alphabet of the language, with some language-specific tweaks described below, whereas the regex model focuses on documented and generalisable error types for the language in question.

6.1. Baseline error models for South and North Sámi

The baseline error models for North and South Sámi are the ones used in production[15]. They are both built following the same structure, and as such the models will be described only once. A general description of the production error model can be found online[16].

The starting point is a Levenshtein edit distance (Levenshtein et al. 1966) error model based on the alphabets of the language, with an editing distance of two. It is possible to adjust the weight of specific edits in the edit distance 2 error model. Adjacent swaps are not enabled by default (they are computationally quite expensive in the present implementation).

Parallel to the default Levenshtein error model, there is a separate set of string edits, handwritten based on identified and frequent error patterns in the languages. The string edits are single FST operations, although each string can be arbitrarily long, thus allowing for much more complex edits than the default model. The string edits are applied as many timed as the default error model, that is, up to twice for both North and South Sámi.

Another extension to the default model is one of suffix edits. That is, a simple transducer mapping input strings to output strings, as the string edits described above, but now restricted to the end of the word. As described above, errors in suffixes are relatively common in especially South Sámi, and this module is meant to target such errors.

Finally, there is a whole-word string replacement module, but that one is utilized very rarely, and does not impact the performance very much. It is also applied to the new regex models described below, mainly because it would be more work to avoid using it.

For Estonian, the regex model is the first one implemented in FST. It is based on the earlier work by Filosoft; no earlier baseline models have been developed for Estonian.

6.2. Rule-based error models

The *regular expressions* (regexes) are grouped according to our assumptions about the nature and likelihood of different types of spelling errors. Also, although guided by the principle that when ranking, one should prefer suggestions with fewer modifications, ours is not based directly on Levenshtein distance. The reasoning is that when calculating the amount of difference between two words, one should view them not as mere symbol strings, but as the traces of a series of mental and physical actions. A change in one action may result in multiple changes in the letter sequence, but it should still be counted as one error.

- Keyboard and orthography (mis)matches. In addition to the Latin letters that form the core of the alphabet, languages typically need some (usually accented) modifications of some of these letters, corresponding to the phones not covered by the core alphabet. These accented letters tend to be positioned in the periphery of the standard keyboard, and/or need key combinations to be used for appearing in the text. It is to be expected that such letters also tend to be mistyped. Also, an accent on a letter may indicate a minor pronunciation subtlety which the speakers need not pay much attention to, so mixing similarly looking and sounding letters would be easy.

  For Estonian, the misspelling list indicates that in case the keyboard does not provide a convenient way to type the accented letters, users may come up with an alternative orthography, e.g. use *sh* or *s^* instead of the correct *š*. If this is the case, then one may expect unlimited substitutions of this kind in a wordform (in addition to other errors). Nordic letters that are not part of the Sámi alphabets, and *á* which is notoriously difficult for North Sámi writers to use correctly, also belong to this class of errors. Correcting them is weighted lightly, and the number of such edit operations is not limited.

[15] https://divvun.no
[16] https://giellalt.uit.no/proof/TheSpellerErrorModel.html

- Keyboard errors, like transposition of letters and repetition of letter sequences, happen so likely in Estonian that regexes for them are needed, while in Sámi, they are highly unlikely. Encoding a context-dependent regex (like one that is needed for repetition) is very costly in terms of FST memory, thus they are not used in the Sámi FSTs.
- Morphology errors, i.e. violating the rules that govern how a word is modified when it is inflected or compounded. These errors are corrected by highly specialised regexes containing string pairs, e.g. a pair of inflectional suffixes.
- Orthography, i.e. the convention of writing phones and their combinations. Letters and combinations that sound similar, like *i* and *j*, belong to this group. For Estonian, the set of orthography-related regexes is smaller than for the Sámi languages, reflecting the proportion of this type of errors in the misspelling list. Also, it is rather common for a Sámi word to contain more than one orthography error (as defined currently); it is possible that a better understanding of the errors will allow us to see in the future how they might really be the manifestation of single errors in the mental process of the writer.

There are different ways to write and combine regexes to yield an FST that converts an input string into another. It is common knowledge among programmers that every existing program can be turned into one that either 1) is smaller when compiled, 2) runs faster, or 3) is more readable, but it is not possible to achieve all these three goals simultaneously. The same is true for FST's. It is well known that an obvious and simple (for the human eye) set of regular expressions may well result in a huge transducer. Aiming at a smaller transducer, one must note that as a rule of thumb, a simpler and smaller transducer puts fewer restrictions on the language it accepts, in other words, the set of possible string pairs passing through a simpler typo modification transducer is larger, thus resulting in more time the speller FST has to spend checking them. Consequently, the number of possible modifications must be controlled, and this forces one to either complicate the regexes or allow the transducer to grow in size.

Appendix A presents a selected set of regex examples showing solutions to some specific problems.

6.3. Machine learned error models

Error modelling in the neural framework is based on imagining the problem as a question similar to machine translation, or just a sequence to sequence character string mapping. Instead of e.g. learning a mapping of e.g. English to French we make the model learn the mapping of misspelled to correct word-forms, and instead of a sentence of words as a context, we have the letters in a word-form. The idea is that if we have enough such mappings, the neural model will learn to translate the misspelled strings into correctly spelt ones, as long as the word list is representative of the errors that are being made. The error correction models that are learnt are character-based, so in principle a representative sample should have some examples of each substitution, deletion and insertion in various enough contexts, so it will learn to make them exactly in the places needed. As is usual with machine learning, the modelling is data-hungry, which means that for ideal usable models we need hundreds of thousands of examples, something that we cannot easily deliver with a low-resource languages. However, in recent years the requirement of the amount of data has been getting smaller, which has made it more plausible to perform these experiments in real low-resource settings.

7. Evaluation

The data in Table 2 gives an overview of the performance of the various error models, for a number of parameters:

- **Spelling error list size:** Number of spelling errors in test corpus for rule-based model, number of training samples / validation + testing for neural network.
- **Average position of correct suggestion:** Ideally this should be 1, ie the correct suggestion is always on top.
- **Average number of suggestions per misspelling:** Ideally this should also be 1, ie there should be no other suggestions than the correct one. That is, the higher the number, the higher the noise level.
- **Top 1/5/all positions:** How many of the misspellings have a correct suggestion in the top position, among the top 5 suggestions, or anywhere among the suggestions
- **No suggestion:** How many of the misspellings have no suggestions; neural models will generally always generate suggestions
- **Only bad suggestions:** How many of the misspellings get only wrong suggestions?

	Estonian		North Sámi			South Sámi		
	RGX	ML	BL	RGX	ML	BL	RGX	ML
Spelling error list size, in thousands	3.0	2.4/0.6		8.5	10.0/1.1		1.1	6.6/0.8
Average position of correct suggestion	1.31	1.97	1.33	1.36	1.99	1.45	1.37	1.49
Average number of suggestions per typo	5.47	6.0	4.00	7.80	6.30	9.30	7.43	6.06
Top 1 positions, %	76.81	13.35	65.03	75.92	46.64	71.32	69.06	34.32
Top 5 positions, %	93.71	28.01	77.53	89.68	63.82	89.43	84.23	46.26
All positions, %	94.46	28.01	78.55	91.30	64.62	91.16	86.05	47.01
No suggestion, %	1.94	0	8.99	2.09	0	1.04	3.55	0
Only bad suggestions, %	3.60	71.98	12.46	6.60	35.38	7.80	10.40	52.99
Speed, words/second	11.05	85.51	31.77	69.34	17.09	14.12	35.69	38.66
FST/ NMT error model size, Mb	13	38	30	31	38	7.9	17	38

Table 2: FST / machine learnt performance; BL = baseline, RGX = handmade regex, ML = machine learnt. ML spelling error list size is specified as training data size / test & evaluation data size.

- **Speed, words/second:** This number is relative, and is provided only to compare between the models and languages. The speed tests were run on the same computer, with as similar conditions as possible; the neural models used a single GPU core and the FST a single CPU core[17]
- **Error model size in megabytes:** FST size is provided in the table to compare the models. The FST size is directly proportional to the use of regexes that consider longer context, like when checking letter pair, triple etc. repetitions, or counting the allowed number of edit operations. The neural model size is the size of the neural network and dependent on the hyperparameters used.

The performance numbers are not directly comparable between languages, due to the different nature of source texts of the misspellings: the orthographic conventions, text creation agents (fast-typing journalists vs a heterogeneous group of Sámi writers), and age of literacy and literary traditions. Keeping these differences in mind, there are still interesting observations to be made.

First, the machine learning models are not able to compete with any of the FST models, not the baseline model, and not the handwritten regex error models. North Sámi had the largest training material, a bit over 8000 typo-correction pairs used for training, and that is clearly not enough to achieve a useful error model. Having established that as a fact is very helpful when guiding future work on minority and indigenous languages. This provides strong evidence that the GiellaLT philosophy is correct: for languages with little to none electronic resources, rule-based is the only option, and we clearly establish that this includes the error models in spelling checkers. If machine learning methods are not currently viable for North Sámi, the language in this experiment with the largest resources, then clearly it will not work for lesser resourced languages.

Upon closer inspection of the ML model, it also became clear that it was mostly simple edit distance one errors that got correct suggestions, which means that it can't contribute meaningfully in a hybrid setup either — the errors it can correct are errors that the rule based model has no problem correcting.

Second, building hand-tuned regular expressions is an exercise worth undertaking, but requires thorough analysis of the error landscape. The Estonian error model performs really well, and the North Sámi handwritten regex model (labelled RGX in Table 2) makes a major leap forward in both recall and precision, making 10+ percentage jumps compared to the baseline. Admittedly, the baseline error model was not that good in the first place, but with the handwritten regex, North Sámi is close behind the Estonian model in performance.

South Sámi, on the other hand, had a good baseline error model to begin with, and with the most complex and varying error typology, it was harder to write a regex that would improve upon the baseline. The regex model is still not very far behind, and with some more analysis and fine-tuning it should be possible to surpass the baseline. The South Sámi spelling error lists have a large portion of errors due to pure orthographic conventions, mixing **ø* for *ö*, and to a

[17] an intel Core i7 CPU and an nVidia Quadro T1000

less degree **ä* for *æ*. Since these errors are frequent, and at the same time very easy to correct, any speller would be expected to perform relatively well. To really improve the South Sámi speller one would have to focus on the remaining classes of misspellings, but as shown above, those are quite heterogeneous. And finally, the misspelling data available to regex development and testing was quite small, and that may influence the numbers. We used another data source for ML training and testing, hence the very different dataset sizes for ML and RGX.

The reasons for the difference between the North and South Sámi baseline models can probably be attributed to several factors. One possible factor is the differences in orthographic principles, where North Sámi often uses accented letters and is thus increasing the size of the alphabet, where South Sámi uses consonant clusters to express the same sound. Another factor is the very different morphophonologies: North Sámi has a very rich and complex consonant gradation system, South Sámi does not have consonant gradation, and South Sámi on the other hand has an elaborate umlaut system, where North Sámi has a very limited set of vowel alternations. But these are just educated guesses on what the cause of the difference could be. A more thorough explanation would require a separate study, and is outside the scope of this article.

We did not use a list of unseen misspellings when evaluating the BL and RGX models. We are aware that when writing regexes, it is possible to come up with specific ones for a selected set of words, i.e. end up with overfitting to the data. We tried to avoid this, and interested readers might check the source code[18]. However, in case of a corpus with different parameters than we had (e.g. a different type of text, different types of text producers), we expect the precision figures to be different.

While having a real held out error corpus for evaluation would be ideal, it is not easily attainable in the context of lesser resourced languages, where we often want to use all available data during development phase for practical reasons.

8. Discussions and speculations, conclusion

Given that for all three languages, one can take an FST speller, baseline or new RGX, that has a recall of over 90%, the main remaining task is to improve precision. That can be achieved in two ways: by even more targeted hand-crafted regexes, or by looking at the context of the error word to either filter or rerank the suggestions. Handcrafting could target the hypothesis that some co-occurring errors are actually inter-dependent (and the co-occurring problem of long-distance dependencies), and/or fine tune the weights and thus the ordering of the suggestions. Since most speller API's do not give any context information, one should try to improve the ordering independently of the context. This is an area for future research.

On the other hand, if the context is available, e.g. via a grammar checker API, it should be possible to filter or promote specific suggestions based on the syntactic context. One such example is Pirinen et al. (2012), using a POS trigram model to promote suggestions matching the trigram model. Another approach is to use the full sentence as the context, in combination with syntactic disambiguation and parsing, for example using the VislCG3 formalism. Examples of such systems are Bick (2006) and Wiechetek et al. (2019), the last one being implemented within the GiellaLT framework, and as such principally available to all languages. That would thus be the next logical step.

For the machine learning setup, it seems that the limitation posed by the amount of training data we have available is still too dire for our use case. While this can be improved by approaches such as automated generation of synthetic misspelling lists, it is limited by the fact that the error generation algorithm should be representative of the errors that real users are likely to make; merely generating statistical noise using a Levenshtein style algorithm will only lead to a neural model that is equal to rule-based Levenshtein corrector but heavier. On the other hand, if we know enough of the nature of the real-world errors to devise an algorithm to generate a representative synthetic misspelling list, we already have an algorithm that can also solve those errors, perhaps more efficiently than the network, thus the main positive side of the neural model may be in the added robustness. All in all, it is good to know that chasing machine learning ghosts can now be put aside, as we have demonstrated that there are better alternatives with a lower environmental impact (see Strubell et al. (2019) for an evaluation of the environmental impact of modern machine learning).

[18] https://github.com/giellalt/lang-sma/, https://github.com/giellalt/lang-sme/, https://github.com/giellalt/lang-est-x-utee/

Acknowledgments

The work was partly supported by the Centre of Excellence in Estonian Studies (CEES, TK-145). The computations for ML model building were performed on resources provided by UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway.

References

Antonsen, Lene. 2013. Čállinmeattáhusaid guorran [english summary: Tracking misspellings]. *Sámi diedalaš áigecála* 2/2013: 7–32.

Arwidsson, Adolf Ivar. 1822. Ueber die ehstniche orthographie. won einem finnländer. *Beiträge zur genauern Kenntniss der ehstnischen Sprache. Funfzehntes Heft* pp. 124–130.

Beeksma, Merijn, Maarten Van Gompel, Florian Kunneman, Louis Onrust, Bouke Regnerus, Dennis Vinke, Eduardo Brito, Christian Bauckhage, and Rafet Sifa. 2018. Detecting and correcting spelling errors in high-quality dutch wikipedia text. *Computational Linguistics in the Netherlands Journal* 8: 122–137.

Beesley, Kenneth R and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford* .

Bergsland, Knut. 1994. *Sydsamisk grammatikk*. Davvi Girji o. s., Karasjok.

Bick, Eckhard. 2006. A constraint grammar based spellchecker for danish with a special focus on dyslexics.

Bollmann, Marcel and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 131–139. The COLING 2016 Organizing Committee, Osaka, Japan.

Bull, Ella Holm and Knut Bergsland. 1974. *Lohkede saemien. Sørsamisk lesebok*. Grunnskolerådet, Kirke- og undervisningsdepartementet: Universitetsforlaget, Oslo.

Erelt, Mati, Tiiu Erelt, and Kristiina Ross. 2007. *Eesti keele käsiraamat*. EKI, Tallinn.

Flor, Michael, Michael Fried, and Alla Rozovskaya. 2019. A benchmark corpus of english misspellings and a minimally-supervised model for spelling correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 76–86. https://doi.org/10.18653/v1/W19-4407.

Flor, Michael, Yoko Futagi, Melissa Lopez, and Matthew Mulholland. 2015. Patterns of misspellings in L2 and L1 English: a view from the ETS Spelling Corpus. *Bergen Language and Linguistics Studies* 6. https://doi.org/10.15845/bells.v6i0.811.

Gaup, Børre, Sjur Moshagen, Thomas Omma, Maaren Palismaa, Tomi Pieski, and Trond Trosterud. 2005. From xerox to aspell: A first prototype of a north sámi speller based on twol technology. In *International Workshop on Finite-State Methods and Natural Language Processing*, pp. 306–307. Springer. https://doi.org/10.1007/11780885_37.

Giellatekno and Divvun. 2021. SIKOR UiT Norges arktiske universitets og det norske Sametingets samiske tekstsamling, versjon 01.10.2021.

Hládek, Daniel, Ján Staš, and Matúš Pleva. 2020. Survey of automatic spelling correction. *Electronics* 9 10: 1670. https://doi.org/10.3390/electronics9101670.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9 8: 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Kask, Arnold. 1970. *Eesti kirjakeele ajaloost*. Tartu Riiklik Ülikool, Tartu.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72. Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/P17-4012.

Kukich, Karen. 1992. Techniques for automatically correcting words in text. *Acm Computing Surveys (CSUR)* 24 4: 377–439. https://doi.org/10.1145/146370.146380.

Levenshtein, Vladimir I et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, vol. 10, pp. 707–710. Soviet Union.

Li, Xiangci, Hairong Liu, and Liang Huang. 2020. Context-aware stand-alone neural spelling correction. *arXiv preprint arXiv:2011.06642* https://doi.org/10.18653/v1/2020.findings-emnlp.37.

Magga, Ole Henrik and Lajla Mattsson Magga. 2012. *Sørsamisk grammatikk*. Davvi Girji, Karasjok.

Moshagen, Sjur, Tommi A Pirinen, and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pp. 343–352.

Moshagen, Sjur N. and Trond Trosterud. 2005. Samisk språkteknologi. In *Nordisk Sprogteknologi 2004: Aarbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004*, edited by H. Holmboe, pp. 57–62. Museum Tusculanums Forlag, København.

Nickel, Klaus Peter and Pekka Sammallahti. 2011. *Nordsamisk grammatikk*. Davvi Girji, Karasjok, 2. hapmi = utgave, 1.deaddileapmi = opplag edn.

Pirinen, Flammie, Krister Lindén, et al. 2014. State-of-the-art in weighted finite-state spell-checking. In *Computational Linguistics and Intelligent Text Processing 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II*.

Pirinen, Tommi, Miikka Silfverberg, and Krister Linden. 2012. Improving finite-state spell-checker suggestions with part of speech n-grams. In *Computational Linguistics and Intelligent Text Processing*, edited by Alexander Gelbukh. International Conference on Intelligent Text Processing and Computational Linguistics ; Conference date: 11-03-2012 Through 17-03-2012.

Pirinen, Tommi A and Sam Hardwick. 2012. Effect of language and error models on efficiency of finite-state spell-checking and correction. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, edited by Iñaki Alegria and Mans Hulden, pp. 1–8. The Association for Computational Linguistics, United States. International Workshop on Finite State Methods and Natural Language Processing ; Conference date: 23-07-2012 Through 25-07-2012.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243* https://doi.org/10.18653/v1/P19-1355.

Trosterud, Trond and Linda Wiechetek. 2007. Disambiguering av homonymi i nord- og lulesamisk. In *Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007*, edited by Ante Aikio and Jussi Ylikoski, Suomalais-Ugrilaisen Seuran Toimituksia 253, pp. 347–354. Suomalais-Ugrilainen Seura, Helsinki.

Trón, Viktor, Andras Kornai, György Gyepesi, László Németh, and Péter Halácsy. 2005. Hunmorph: Open source word analysis. In *Proceedings of the Workshop on Software. Association for Computational Linguistics*, pp. 77–85. https://doi.org/10.3115/1626315.1626321.

Wiechetek, Linda, Sjur Nørstebø Moshagen, and Kevin Brubeck Unhammer. 2019. Seeing more than whitespace — tokenisation and disambiguation in a North Sámi grammar checker. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pp. 46–55. Association for Computational Linguistics, Honolulu. https://doi.org/10.33011/computel.v1i.403.

A. Appendix. Some tips and tricks for FST

Below is a selected set of regex examples showing solutions to some specific problems.[19]

A.1. Transposition and permutation

Hitting right keys in a wrong order may result in letter transpositions, e.g. Estonian **blianss* instead of intended *bilanss*, and permutations, e.g. **skepulatsioon* instead of intended *spekulatsioon*. The task of a regex is to re-order a few letters.

Transposition may be encoded as a set of pairs of all letters of the alphabet being swapped, e.g. for adjacent letters:

```
[{ab} -> {ba}] | [{ba} -> {ab}] | [{ac} -> {ca}] | [{ca} -> {ac}] |
...
[{yz} -> {zy}] | [{zy} -> {yz}]
```

However, it can also be modelled as a process where a letter disappears from one side of some other letter, and appears on the other side, that is, becomes zero and emerges from zero. (The other letter - any letter, for that matter - is expressed by a ?-mark, denoting "any symbol".) The resulting FST will also transform a pair of identical letters to the same pair, but in our modification-plus-checking workflow this makes no real harm.

```
# transposition 12 -> 21
[a:0 ? 0:a] | [b:0 ? 0:b] |
...
[z:0 ? 0:z]
```

This way of expressing transposition makes a smaller transducer than would be the alternative with explicitly listed pairwise expressions, because using the ?-mark imposes fewer restrictions on the language than explicitly listing the transposition pairs.

In addition, defining a transposition via a letter disappearing and appearing, it makes one notice that transposition error is just a special case of permutation errors, i.e. transposition is permutation of adjacent letters. For example, the Estonian misspelling list contains a typo **proessorf*, where the error is a permutation *fessor -> essorf*. Below is an example of expressing permutation error correction.

```
# permutation 123 -> 312
[0:a ? ? a:0] | [0:b ? ? b:0] |
...
[0:z ? ? z:0]
```

A.2. Repetition

It may happen that a part of a word is mistakenly re-typed, e.g. Estonian **minimimaalne* instead of intended *minimaalne*, and the task is to delete this repeated part.

Repetition-expressing regexes are notorious for blowing the transducer's size up. In order to alleviate that, one may use the ?-mark again, under-specifying the context and thus arriving at a smaller compiled transducer.

```
# abab -> ab
[
a (->) "<COR>" || _ ? a ,,
b (->) "<COR>" || _ ? b ,,
...
z (->) "<COR>" || _ ? z
]
.o.
[ [? - "<COR>"]* "<COR>" "<COR>" [? - "<COR>"]* ]
```

[19] We use the Xerox regular expression notation, c.f.Beesley and Karttunen (2003)

The regex above ultimately deletes the first letter pair from a sequence of two similar pairs. First, any letter is optionally substituted with a `<COR>` tag, if the same letter also appears as next-to-next. This transducer is then composed with another one, which requires exactly two adjacent `<COR>` tags to be in the resulting string. This ensures that there were exactly two adjacent letters that were substituted, meaning that there was a repetition pattern like `abab` in the wordform. Once `<COR>` tags are removed, the removal of two repeated letters is finished.

A.3. Regex as linguistic abstraction

Finally, below is an example on how to express a linguistic abstraction in a regex shorthand. Sámi writers tend to be confused on how to write similarly sounding phones. For example, *k-sound* may be written in North Sámi as *g*, *gg*, *k*, *kk*, or *hk*, and the list of misspellings contains many word pairs where these letters are confused, e.g. **geatgi - geatki*, **Mákkarávjui - Máhkarávjui*, **sámedikkeválggas - sámediggeválggas*, **Johtolagii - Johtolahkii*, **Sámedigge - Sámedikke*, **ohkiin - ogiin*.

The task is to substitute a wrong letter sequence with a correct one, while keeping the intended sound sequence. It is convenient to model this orthography-related uncertainty via confusion sets. Below is an example of expressing a confusion set of 5 letter combinations that are used for writing down *k-sound*:

```
[ [ { g } | { gg } | { k } | { kk } | { hk } ] : [ { g } | { gg } | { k } | { kk } | { hk } ] ]
```

The expression means that any of these combinations may be substituted for any other. The regex also redundantly substitutes a combination for itself, resulting in a ”modification” that is the same as the original; such modifications get discarded by the speller module, so they make no harm except waste a little extra time.

Kantasaamen sensiivisen *-*kšǥ*-johtimen kehityksestä ja edustuksesta nykysaamessa

Eino Koponen ja Juha Kuokkala
Oulun yliopisto, Helsingin yliopisto

Abstract

The article discusses Saami censive verbs containing the suffixal element *-š-*, such as North Saami *guhkášit* ‘consider (too) long’ (of *guhkki* ‘long’). The occurrence of individual derivatives and derivational subtypes across the Saami languages are studied on the basis of extensive dictionary data, and the outlines of the historical development of the derivational type are sketched. Considering the Inari Saami verbs of type *viššâlšukšâđ* ‘consider diligent’ and data from past centuries, it is argued that the derivational type goes back to Proto-Saami *-*kšǥ*-, which, in turn, is a loan suffix from Finnic (cf. Finnish *kummeksua* ‘find something odd’ ← *kumma* ‘odd’, *halveksia* ‘despise’ ← *halpa* ‘cheap’).

Keywords: Sámi languages, verbal derivation, historical morphology

1. Johdannoksi

Tämä artikkeli on toinen osa kolmeosaiseksi tarkoitetusta artikkelisarjasta, jossa tarkastelun lähtökohtana ovat infinitiivissä kolmitavuiset *-šit*-loppuiset pohjoissaamen verbit ja niiden vastineet muissa saamelaiskielissä. Nielsenin sanakirjan pohjalta laaditussa pohjoissaamen käänteissanaluettelossa (Sammallahti 2002: 248–249) tällaisia verbejä on n. 290. Sekä semanttisesti että pohjoissaamen ulkopuolella myös äänteellisesti selvimmin muista erottuva ryhmä (99 kpl) ovat kantasaamen *-*hčǥ*-johtimella muodostetut verbikantaiset frekventatiiviverbit, joita käsittelimme artikkelisarjan ensimmäisessä osassa (Koponen & Kuokkala 2021).

Tässä osassa otamme tarkasteltavaksi kantasaamen *-*kšǥ*-johtimella muodostetut adjektiivikantaiset sensiivijohdokset, joiden semanttinen tuntomerkki on se, että ne merkitsevät verbin objektin pitämistä sellaisena, että sillä on (usein epätoivotussa määrin) kantasanan ilmaisema ominaisuus.[1] Tyypillisiä sensiiviverbejä ovat esim. pohjoissaamen *guhkášit* ’pitää (liian) pitkänä’ (vrt. *guhkki* ’pitkä’) ja *oanášit* ’pitää (liian) lyhyenä’ (vrt. *oanehas* ’lyhyt’). Sensiivijohdosten ja muiden (nominikantaisten tai kantasanattomien) *-šit*-verbien raja ei ole aivan yksiselitteinen. Palaamme tähän kysymykseen artikkelisarjamme kolmannessa osassa, jossa käsiteltävänä ovat sensiivijohdoksiin ja frekventatiivijohdoksiin kuulumattomat *-šit*-verbit. Sensiiviverbeiksi tulkitsemiamme verbejä on yllä mainitussa pohjoissaamen sanaluettelossa 145 kpl.

Tutkimusaineistomme on peräisin keskeisesti samoista sanakirjalähteistä kuin edellisessä osassa mutta eräin täydennyksin.[2] Erityisesti kildininsaamesta olemme nyt ottaneet mukaan myös Rimma Kuručin,

[1] Käytämme saamentutkimuksessa suhteellisen vakiintunutta termiä *sensiivi(nen)*, englanniksi *censive* nimenomaan tässä asussa pitäen kantana latinan verbiä *cēnseō* ’arvioida’. Vastaavasti *c*:llistä kirjoitusasua *censitiv* käyttävät Nielsen (1979 [1926] § 309; norjaksi) ja Ruong (1943: 141, 170; saksaksi), Itkonen (1980: 27) saksaksi *censiv*. Eräissä uudemmissa lähteissä, kuten Sammallahti 1998 (s. 93), esiintyy sen sijaan englanninkielinen asu *sensive* (vrt. Nickel 1990: 292 norjaksi *sensiv*), joka yhdistyy latinan verbiin *sēntiō* ’tuntea, kokea, aistia’. Johdostyypin semantiikka liittyy kuitenkin ennen kaikkea (kriittiseen) arviointiin pikemmin kuin aistimiseen. VISK (§ 350) käyttääkin suomen vastaavista *-ksu*-verbeistä kuten *oudoksua* nimitystä *suhtautumisjohdos* eli *sensiivijohdos*.

[2] Koltansaamessa Ve′rdd-tietokannan korvaa sen pohjalta laadittu uusi Suomi–koltansaame-sanakirja (Moshnikoff & Moshnikoff 2021). Täydentävinä lähteinä on mukaan otettu teokset Barruk 2018 (uumajansaame), Halász 1896 (piitimensaame), Grundström 1946–1954 (luulajansaame) ja Sammallahti 2021 (pohjoissaame). Aineiston kokoamiseen on käytetty mahdollisuuksien mukaan sanakirjojen sähköisiä tai digitoituja versioita. Kaikki aineisto-

 Nordlyd 46.1: 141–158, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, toimittajat Lene Antonsen, Sjur Nørstebø Moshagen ja Øystein A. Vangsnes. Kustantaja UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd
https://doi.org/10.7557/12.6446

Nina Afanasjevan et al. vuonna 1985 julkaiseman teoksen Saamsko-russkij slovar (SRS). Tämä sanakirja olisi ollut syytä ottaa huomioon jo ensimmäisessä osassa, koska sen n. 150 kildininsaamelaista *-*hčę*-frekventatiivijohdosta vahvistavat havaintoamme frekventatiivijohtimen esiintymisestä kildininsaamessa koltansaamen tapaan myös *-*ō*- ja *-*ę*-vartaloihin liittyneenä ja antavat lisävalaistusta avoimeksi jääneeseen kysymykseen *-*ō*-vartaloisten verbien johdinta edeltäneen vokaalin laadusta. Tässä aineistossa ensimmäisen tavun vokaali on johdonmukaisesti korkea, mikä viittaa toisen tavun pyöreän vokaalin olleen (johtimen *-*ę*-vokaalin vaikutuksesta) pohjoissaamen tapaan *-*u.*

Nyt tarkasteltavan *-*kšę*-sensiivijohtimen sisältäviä verbejä on käytössämme olleista muiden saamelaiskielten aineistoista löytynyt vaihtelevia määriä, ja on epäselvää, missä määrin lähteiden esiintymämäärät kuvastavat niiden todellista yleisyyttä eri saamelaiskielissä ja missä määrin vain aineistojen suppeutta sekä sanakirjan laatijan suurta tai vähäistä kiinnostusta tätä johdostyyppiä kohtaan. Ei ole yllättävää, ettei puutteellisimmin dokumentoidusta turjansaamesta ole tiedossamme yhtään kolmitavuista[3] sensiivijohdosta ja uumajansaamestakin vain kolme, eikä ehkä sekään, että KKLS:sta tällaisia on löytynyt kaikkiaan vain 23 (15 kildininsaamesta ja 8 koltansaamesta[4]). Yllättävää sen sijaan on, että kildininsaamea edustavassa SRS:ssa niitä on 149, ja yllättävänä voitaneen pitää sitäkin, että Lagercrantzin sanakirjasta (LW) sensiivijohdoksia on löytynyt vain muutamia.[5]

Artikkelin luvussa 2 tarkastelemme nykyisten saamelaiskielten sanakirjoista löytämiemme 290 kolmitavuisen sensiivijohdoksen levikkiä eli tietyn johdoksen esiintymistä useammassa tai vain yhdessä kielessä. Luku 3 on omistettu runsasjohdoksisen ja läntisemmistä kielistä johtosuhteiltaan poikkeavan kildininsaamen yhteen kieleen rajoittuvien johdosten semantiikan ja kantasanan morfologian analysoinnille. Luvussa 4 laajennamme tarkastelukulmaa nelitavuisiin sensiivijohdoksiin, ja luvussa 5 käsittelemme johdostyypin derivotaksia eli sitä, millaisiin kantasanoihin ja niiden vartaloallomorfeihin johdin liittyy. Luvussa 6 hahmottelemme sensiivijohdosten eri tyyppien syntytapaa ja kehitystä, ja luvussa 7 kokoamme yhteen tutkimuksen tuloksia.

2. Kolmitavuisten johdosten levikistä

Nykyisten saamelaiskielten aineistomme käsittää 73 useammasta kuin yhdestä saamelaiskielestä ja 217 vain yhdestä kielestä löytynyttä kolmitavuista *-*kšę*-sensiivijohdosta. Useammassa kielessä esiintyvät johdokset on esitetty liitteessä 1. Eniten muiden kielten kanssa yhteisiä johdoksia on pohjois- (69), inarin- (54) ja kildininsaamessa (31). Vain yhdestä kielestä tavattuja sensiivijohdoksia on eniten kildininsaamessa (123) ja pohjoissaamessa (80). Esiintymämäärät eri kielissä selviävät taulukosta 1. Pelkästään kildininsaamessa esiintyvät johdokset on listattu liitteessä 2 ja muiden saamelaiskielten omat johdokset vastaavasti liitteessä 3.

Kieli	Et	U	Pi	Lu	Po	In	Ko	Ki
Useammassa kielessä esiintyviä	8	3	13	16	69	54	9	31
Vain yhdessä kielessä esiintyviä	2	-	1	4	80	8	-	123
Sensiivijohdoksia yhteensä	10	3	14	19	149	62	9	154

Taulukko 1. Kolmitavuisten sensiivijohdosten määrät eri saamelaiskielten aineistoissa.

lähteet lyhenteineen on lueteltu artikkelin lopussa. Tutkimusaineisto on keskeisiltä osiltaan koottu liitteisiin 1–4, jotka ovat saatavissa myös erillisinä tiedostoina Zenodo-palvelussa osoitteessa https://doi.org/10.5281/zenodo.6471592 .

[3] Kolmitavuisiksi nimitämme johdoksia, joiden johdinaines aloittaa sanan (alkuperäisen) 3. tavun ja nelitavuisiksi vastaavasti johdoksia, joiden johdinaines sijaitsee 4. tavussa.

[4] Koltansaamen johdoksista 7 on kildininsaamen johdosten vastineita. Ainoa KKLS:n vain koltansaamesta tuntema kolmitavuinen sensiivijohdos on *ååudšed* 'oudoksua', ja siitäkin annetaan vain inkoatiivijohdoksen muoto *ååudšeškuõ'đi* (s. 325).

[5] Ilmaus "on löytynyt" on sekä KKLS:n että LW:n osalta sikäli paikallaan, että näiden tarkekirjoituksella kirjoitettujen monikielisten sanakirjojen formaatti ei mahdollista puheena olevaan johdostyyppiin kuuluvien sanojen kattavaa tekstihakua pdf-muodosta.

Saamelaiskielet voidaan jakaa (osaksi käytännöllisin, osaksi myös teoreettisin perustein[6]) kolmeen ryhmään, joista eteläryhmän muodostavat eteläsaame, uumajansaame ja piitimensaame ja itäryhmän koltan-, kildinin- ja turjansaame. Näiden väliin jääviä kolmea kieltä (luulajansaame, pohjoissaame ja inarinsaame) voidaan nimittää keskiryhmäksi.

Aineistomme 73 verbistä 9:n levikki ulottuu niin etelä-, keski- kuin itäryhmäänkin, 8:n levikki ulottuu etelä- ja keskiryhmään, mutta ei itäryhmään, ja 24:n keski- ja itäryhmään, mutta ei eteläryhmään. Levikiltään vain keskiryhmään rajoittuvia verbejä on 32. Aineistossamme ei ole sensiivijohdoksia, joiden levikki rajoittuisi kahteen tai kolmeen eteläryhmän tai itäryhmän kieleen, eikä asia sanottavasti muutu, vaikka eteläryhmään luettaisiin kuuluvaksi vielä luulajansaame tai itäryhmään inarinsaame. Kumpaankin tulisi näin yksi verbi, itäryhmään In *rođášiđ* ’pitää rumana’ ja eteläryhmään Lu *nuohkahit* ’pitää riittävänä’.

3. Kildininsaamen johdoksista ja niiden kantasanoista

Kildininsaamen aineisto on peräisin kahdesta sanakirjasta, joista toinen on Itkonen 1958 (= KKLS) ja toinen Kuruč & al. 1985 (= SRS). Ensin mainittu lähde sisältää aineistoa kildininsaamen ohella muistakin koltan- ja kuolansaamen varieteeteista, jälkimmäinen taas edustaa kildininsaameen perustuvaa venäjänsaamen kirjakieltä. Aineistoon kuuluu 31 sensiivijohdosta, joille on löytynyt vastine vähintään yhdestä muustakin saamelaiskielestä, ja 123 johdosta, joille ei ole löytynyt vastineita kildininsaamen ulkopuolelta.

Ne kildininsaamen sensiivijohdokset, joilla on vastineita muissa saamelaiskielissä, on esitetty liitteessä 1 siten, että lähteessä SRS esiintyvät 25 johdosta on kirjoitettu tämän sanakirjan ortografialla ja vain lähteessä KKLS esiintyvät 6 sanaa koltansaamen ortografialla ja koltansaameen ”transponoituina”. Molemmissa lähteissä esiintyvät 9 johdosta on merkitty sanan lopussa olevalla asteriskilla. Lähteessä KKLS esiintyviä johdoksia on siis yhteensä 15 ja lähteessä SRS esiintyviä 148.

Käyttämiemme sanakirjojen valossa vain kildininsaameen rajoittuvat sensiivijohdokset on lueteltu liitteessä 2. Niistä kolme asteriskilla merkittyä esiintyvät molemmissa lähteissä, kaikkien muiden ainoa lähde on SRS. Liitteen 2 sensiivijohdoksista 64 kuuluu tyyppiin, jossa sanueen kantasana on substantiivi ja sensiivijohdoksen välitön kantasana on (ainakin semanttisesti) mainitun substantiivin *-ай*/*яй*-johdos, jonka merkitys on ’sellainen, jossa on runsaasti kantasanan tarkoitetta’. Tällainen on esimerkiksi sensiivijohdos *ёӈӈшэ* ’pitää runsaspuolukkaisena’, jonka välitön kantasana on substantiivista *ёӈӈ* ’puolukka’ johdettu possessiiviadjektiivi *ёӈӈяй* ’runsaspuolukkainen’.[7] Samaan tyyppiin kuuluu lisäksi kaksi sensiivijohdosta, joiden yhteyteen kuuluva possessiiviadjektiivi tunnetaan vain kildininsaamen ulkopuolelta (*лоастшэ* ’pitää runsaslastuisena’ ~ Ko *la'stti* ’lastuinen’, *сӣйшэ* ’pitää visvaisena’ ~ Po

[6] Tässä käyttämämme eroaa perinteisestä äänne- ja muotopiirteisiin perustuvasta ryhmittelystä siten, että viimeksi mainitussa inarinsaame kuuluu itäryhmään ja piitimensaame keskiryhmään, vrt. Rydving 2013.

[7] Nimitämme puheena olevan tyyppisiä saamen adjektiiveja possessiiviadjektiiveiksi, koska ne vastaavat funktioltaan VISK:n possessiivisiksi nimittämiä suomen kielen johdostyyppejä *kalaisa* (§ 285) ja *lehtevä* (§ 290). Johdin palautunee esisaameen, joskin sen historiasta on esitetty erilaisia käsityksiä (ks. Korhonen 1981: 323–324; Sammallahti 1998: 91). Kantasaamen myöhäisvaiheessa johtimen asu on ollut kaksitavuisiin nominivartaloihin liittyessään (vartalovokaali **ę*, **ā*, **u* +) **-ję*. Attribuuttijohtimen **-s* (ja monikon tunnuksen **-k*) edellä johtimen **j* on kadonnut, jolloin sen molemmin puolin olleet vokaalit ovat sulautuneet supistumavokaaleiksi **î*, **â*, **û*. Tämän jälkeen eri vartalotyyppien supistuneet ja supistumattomat muodot ovat vaikuttaneet toisiinsa, ja kehityksen lopputulosta kuvastavat nykykielten asut Po *lusʹsii* ~ *luosʹsái* ’runsaslohinen’, *vuddjii* ~ *vuoddjái* ’rasvainen’, *geađgái* ’kivinen’, *ulʹlui* ’villaisa’, *vuđʹđui* ’laajapohjainen’, vastaavasti In *vuojjii*, *kiäđgáá*, *ulluu*, *vuođđuu*, Ko *luõʹsʹsi*, *vuõʹjʹji*, *ǩeädggai*, *vuâđʹđai*, Ki *лӯссай*, *кӣодкай*, *уллай* (ksa. **luosęję*, **vuojęję*, **keädkāję*, **ulluję*, **vuoduję*), sekä attribuuttimuodot Po *lusʹses* ~ *luosʹsás*, *vuddjes* ~ *vuoddjás*, *geađgás*, In *vuojjiis*, *kiäđgáás*, *ulluus*, Ko *luõʹsʹses*, *vuõʹjʹjes*, *ǩeädggas*, *vuâđʹđas*, Ki *уллъесь* (ksa. **luosîs*, **vuojîs*, **keädkâs*, **ullûs*, **vuoduûs*). SSS ei anna adjektiiveille *ulʹlui*, *vuđʹđui* (eikä ilmeisesti muillekaan tähän tyyppiin kuuluville sanoille) *s*-loppuista adjektiivimuotoa. SRS:ssa on näistä esimerkkisanoista attribuuttimuoto vain yhdellä. KKLS:ssa samanlainen supistumaton attribuuttimuoto kuin mainittu *уллъесь* on myös adjektiivien *ǩeädggai* ja *vuâđʹđai* kildininsaamelaisilla vastineilla; sanan *vuõʹjʹji* vastineen attribuuttimuodosta ei ole yksiselitteisesti pääteltävissä, onko sekin supistumatonta tyyppiä vai vastaako se koltan muotoa *vuõʹjʹjes*. SRS:ssa attribuuttimuotoja on satunnaisesti, ja ne lienevät johdonmukaisesti supistumattomia: *аббьръесь*, *пӣнньесь.*

sieddjái ’visvainen’) sekä kuusi sensiivijohdosta, joiden yhteyteen kuuluvaa possessiiviadjektiivia ei ole kirjattu mistään saamelaiskielestä, vaikka sellainen voidaan sensiivijohdoksen merkityksen perusteella olettaa. Sensiivijohdoksista 34 kuuluu puolestaan tyyppiin, jossa sensiivijohdoksen (ja usein koko sanueen) kantasana on johtimeton tai synkronisesti hämärän johtimen sisältävä adjektiivi.

Osasta kildininsaameen rajoittuvia sensiivijohdoksia on vaikea ratkaista, kumpaan (jos kumpaankaan) tyyppiin ne kuuluvat. Sensiivijohdoksen *сунитшэ* merkitys on SRS:n mukaan ’pitää sellaisena, jossa on paljon sulapaikkoja’ ja kantasana on *сунньт* ’sulapaikka’. KKLS:n mukaan koltansaamen *su'dd* ’sula(paikka)’ kildininsaamelaisine vastineineen voi olla paitsi substantiivi myös adjektiivi, jolloin sillä on attribuuttimuoto *suu'ddes*; sanan turjansaamelaiselle vastineelle KKLS antaa vain adjektiivisen merkityksen. Pohjoissaamessa *suddi* on SSS:n mukaan substantiivi ja adjektiivi on *suttis* (mon. *suddásat*); lisäksi SSS tuntee *-i*-johdoksen *suddái* ’sellainen, jossa on paljon sulapaikkoja’. Kaikilla kolmella sanalla on ILW:n mukaan tarkka vastine inarinsaamessa. Sanoista ensimmäisellä on vastine kaikissa länsisaamelaisissa kielissä (LW § 7078.2; YSS § 1164) ja toisellakin ainakin luulajan-, uumajan- ja eteläsaamessa. Sanueeseen kuuluu myös laajalevikkinen verbi Po *suddat*, Ki *сунитэ* ’sulaa’.

Edellä puheena olleelle sensiivijohdokselle sekä morfologisesti että semanttisesti läheinen on kildininsaamen *муллвшэ*, jonka merkitys on SRS:n mukaan ’pitää runsasvetisenä’ ja kantasana *муллв* ’tulva’ (vastine kaikissa saamelaiskielissä; YSS § 1286). Johdoksen pohjoissaamelaisen vastineen *dulvvášit* merkitys on SSS:n mukaan ’pitää (jokea) liian tulvaisena, pitää tulvaa liian korkeana’. Kantasanan *dulvi* ’tulva’ ohella pohjoissaamen sanueeseen kuuluu adjektiivi *dulvvis* (mon. *dulvásat*) ’tulvainen’, jolla on vastine eteläsaamessa, sekä kantasanan superlatiivijohdos *dulvvimus* ’tulvaisin, korkeimmillaan tulviva’, jolla puolestaan on vastine luulajansaamessa. Koltansaamessa esiintyy adjektiivi *tolvvai* ’tulvainen’, jolla KKLS:n mukaan on vastine turjan- ja kildininsaamessa (SRS ei tällaista tunne).[8]

Kildininsaamen sensiivijohdoksista *кыдтшэ* ’olla sitä mieltä, että kevät on alkamassa’ ja *мāллвшэ* ’olla sitä mieltä, että talvi on pitkä’ ensin mainitulla on Friisin sanakirjassa[9] (norjansaamelainen) vastine *giđašet* (ja *giđašavšat*), jälkimmäisellä pohjoissaamelainen vastine *dálvvášit*, jonka merkitys on Nielsenin mukaan ’find, consider, that it is too wintry (to do something)’. Kantasanan *dálvi* ’talvi’ (vastine kaikissa saamelaiskielissä; YSS § 1223) ohella pohjoissaameen sanueeseen kuuluu adjektiivi *dálvvas* ’talvinen’. Kildininsaamessa esiintyy adjektiivi *māллвай*, jolle SRS antaa merkityksen ’sellainen (esim. seutu), missä on pitkät talvet’ ja KKLS merkityksen ’talvinen (esim. kevät)’. Kantasanalla *кыдт* ’kevät’ tai sen vastineilla (saPo *giđđa*; vastine kaikissa saamelaiskielissä; YSS § 399) ei sanakirjalähteiden valossa ole *dálvvas*, **dálvvis*, **dálvái* -tyyppisiä johdoksia.

Sensiivijohdokset *оаӈӈкшэ* ’olla sitä mieltä, että hanki on jo vahva’ ja *яллшэ* ’pitää tietä tasaiseksi ajettuna’ kuuluvat sanueisiin, joiden levikki sanakirjalähteiden (SRS:n lisäksi KKLS 14, 48, 810, 825) valossa rajoittuu turjan- ja kildininsaameen. Ensin mainittuun sanueeseen kuuluu kildininsaamessa kantasanan *оаӈӈк* ’hanki’ ohella adjektiivijohdos *оаӈӈкай*, jonka merkitys SRS:n mukaan on ’hyvä (hangesta puhuttaessa)’. Myös jälkimmäiseen sanueeseen kuuluu samaa tyyppiä edustava adjektiivi *яллай*, jonka merkitykseksi SRS antaa ’tasaiseksi ajettu (talvitie)’ ja KKLS ’sileä; poljettu (tie lumessa)’. Sanueen kantasana on SRS:n mukaan *ялл* ’tasaiseksi ajettu talvitie’; KKLS tulkitsee tämän asun (edellä mainitun adjektiivin?) attribuuttimuodoksi ja vertaa sanuetta koltan-, kildinin- ja turjansaamessa esiintyvään substantiiviin *jeä'les* ’iljanne, kaljama’. Onkin mahdollista, että *ялл(ай)*-sanueen taustalla on viimeksi mainitun substantiivin *i*-johdos **jeällai* ’iljanteinen’.

Edellä käsitellyille sensiivijohdoksille yhteinen merkityspiirre on, että niiden avulla voidaan ilmaista kriittistä arviota kelin eli kulkuväylän tai sääolojen sopivuudesta. Niiden tyypilliseksi – ja mahdollisesti myös alkuperäiseksi – käyttötavaksi voidaan ajatella tilanteita, jotka ovat tyyppiä ”olen sitä mieltä, että keli

[8] Johtosuhteiltaan tätä muistuttava on sensiivijohdos *мӯллпшэ* ’pitää tylppänä’, jonka kantasana SRS:n mukaan on adjektiivi *мӯлл* tai *мӯлльпай*. Näiden pohjoissaamelaisia vastineita ovat SSS:n mukaan substantiivi *dulpi* ’tylppä esine’ ja adjektiivit *dulpái*, *dulppas* (mon. *dulpasat*, attr. *dulpe-*) ’tylppä’, joista muodoilla *dulpái* ja *dulpe-* on KKLS:n mukaan tarkka vastine koltansaamessa.

[9] Tästä lähteestä samoin kuin muistakin 1900-lukua vanhemmista sanakirjoista olemme ottaneet lähemmin tarkasteltavaksemme vain nelitavuiset sensiivijohdokset.

on ominaisuuden x suhteen (liian) huono”. Sääoloihin rinnastuvat myös vuorokaudenajoista riippuvat valaistusolot. Tähän merkitysryhmään kuuluu kildininsaamen johdos *вя̄ӣйкшэ* ’pitää hämäränä’, jonka kantasanalla *вē̄ӣйк* ’iltahämärä’ on vastineita laajalti itä- ja länsisaamelaisissa kielissä (saPo *veaigi*; YSS § 1369). Muita samaan ryhmään kuuluvia sensiivijohdoksia ovat kildininsaamen *ыӣйшэ* ’pitää ajankohtaa (liian) öisenä (johonkin tarkoitukseen)’ pohjois- ja inarinsaamelaisine vastineineen sekä vain inarinsaamesta kirjattu *piäivášiđ* ’pitää ajankohtaa tai valaistusoloja liian päiväisinä (eli ei riittävän hämärinä)’. Lähinnä tähän ryhmään kuuluu myös sensiivijohdos *мōллшэ* ’pitää (esim. kaupunkia) sellaisena, jossa on paljon valoja’. Johdoksen ja kantasanan (*мōлл* ’tuli’) semanttinen suhde selittynee ainakin osaksi sitä kautta, että venäjän sanan *огонь* merkitys on paitsi ’tuli’ myös ’(lampun) valo’.

Monien sensiivijohdosten yhteinen semanttinen piirre on, että niissä arvioinnin kohteena on tyypillisesti ihminen (tai eläin, esim. poro) jonkin henkisen ominaisuuden suhteen. Sanueeseen kuuluu tällöin tavallisesti ominaisuutta kuvaava johdettu adjektiivi ja sanueen kantasanaksi tulkittavissa oleva substantiivi. Tällaisia ovat kildininsaamen *куарркшэ* ’pitää kerskailevana tai ylpeänä’ (vrt. *куарркэш* ’kerskaileva’, *куаррк* ’kerskailu’) ja *мōԯԯкшэ* ’pitää tolkullisena’ (vrt. *мōԯԯкэш* ’tolkullinen’, *мōԯԯк* ’tolkku, järki’). Adjektiivi *мōԯԯкэш* kuuluu denominaaliseen tai deverbaaliseen johdostyyppiin, joka Korhosen (1981: 324; 3.4.2.6.) mukaan ilmaisee kantasanan tarkoitteeseen liittyvää erittäin vahvaa ominaisuutta, esim. Po *njun'neš* ’herkästi haistava’ (vrt. *njunni* ’nenä’), *dusteš* ’rohkea’ (vrt. *duostat* ’rohjeta’), ja jonka johdin on historiallisesti yhdysperäinen: **-jē* (partisiippijohdin) + **-hčē*. Myös adjektiivi *куарркэш* voi kuulua samaan tyyppiin, mutta avoimeksi jää, olisiko sen alkuperäinen kantasana substantiivi *куаррк* vai nykyisin vain koltansaamesta tunnetun verbin *kuärggad* ’kerskua’ vastine. Inarinsaamesta tunnetaan merkityksessä ’kerskaileva’ adjektiivit *korguš* ja *korguu*. Näistä ensin mainittu näyttäisi olevan kildininsaamen adjektiivin tarkka vastine, jälkimmäinen puolestaan on samaa tyyppiä kuin edellä alaviitteessä 7 mainittu *vuođđuu*, siis ksa. **koarkuję*. Kuitenkin jos adjektiivin *куарркэш* vahva aste on myöhäinen (tai peräti virheellinen), sana voi olla koltansaamen adjektiivin *kuärgač ~ kuärgaš* ’kerskaileva’ vastine; tämä puolestaan kuuluu vartalokonsonantin heikon asteen perusteella deverbaaliseen johdostyyppiin, joka ilmaisee taipumusta kantasanan tarkoittamaan toimintaan, esim. *gološ* ’kylmänarka’ (vrt. *goallut* ’palella’). Historiallisesti tyyppi eroaa edellisestä siten, että jälkimmäisessä johdin **-hčē* liittyy verbivartaloon ilman **-jē*-ainesta. (Korhonen 1981: 324; 3.4.2.7.)

Sensiivijohdos *суһтшэ* ’pitää oikutteluun taipuvaisena’ kuuluu sanueeseen, jonka kantasana on SRS:n mukaan substantiivi *суһт* ’oikku’ ja muita jäseniä *суһтай* ’oikutteluun taipuvainen’ sekä *сүдтнэ* ’oikutella (esim. lapsi)’. Viimeksi mainitun tarkkoja vastineita ovat Po *suhtadit* ’kiukutella’ ja In *sutâdiđ* id.; ne ovat frekventatiivijohdoksia (ksa. **suhtęntę-*) laajalevikkisestä (vaikka SRS:lle tuntemattomasta) kantaverbistä, jota edustavat Po *suhttat* ’suuttua’ ja In *suttâđ* (YSS § 1166). SRS:n *суһт* lienee pohjoissaamessa (ja laajemminkin) tunnetun substantiivin *suhttu* ’suuttumus’ äännevastine eli kantaverbin teonnimi ja *суһтай* joko kantaverbin partisiippi (ksa. **suhtęję̄*) tai teonnimen possessiivijohdos (ksa. **suhtuję*). Lähteen Sammallahti 1998 (s. 91) mukaan pohjoissaamen sanueeseen kuuluisi myös adjektiivi *suhtteš* (< ksa. **suhtęję̄hčē*). Muissa lähteissä (myös SSS ja Nickel & Sammallahti 2011: 636) esiintyy tämän sijaan toiseen tyyppiin kuuluva adjektiivi (ksa. **suhtęhčē*) > *suhteš* ’äkkipikainen’. Adjektiivi *suttiš* esiintyy inarinsaamessakin, mutta emme osaa sanoa, kumpaa tyyppiä se edustaa.

Sekä merkitykseltään että johtosuhteiltaan edellistä muistuttava on sensiivijohdos *нубшэ* ’pitää itsepäisenä’ (vrt. *нубпай* ’itsepäinen’, *нубпъесь* id.), jonka kantasanaksi SRS antaa substantiivin *нубпь* ’itsepäisyys’. KKLS ei tunne tällaista substantiivia mutta kylläkin (s. 909) verbin Ki *nuuppâd* ’olla itsepäinen’ sekä adjektiivin *nu'ppeš* ’itsepäinen’ (< ksa. **nupęję̄hčē*). Adjektiiviin *нубпай* pätee mutatis mutandis sama, mitä edellä on todettu sanasta *суһтай*.

Sensiivijohdoksen *поагшэ* ’pitää uppiniskaisena’ (vrt. *поагкяй* ’uppiniskainen’, *пāгкьесь*, *пāгеч* id.) kantasanaksi SRS antaa substantiivin *пāгк* ’uppiniskaisuus’. KKLS ei tunne tällaista substantiivia, mutta kylläkin (s. 327) verbin *pää'ǩǩed* ’niskuroida’, joka esiintyy (asussa *пāгке*) SRS:ssakin. Substantiivi *пāгк* lienee verbin *пāгке* **-ō-*teonnimi. Sekä SRS:lle että KKLS:lle (s. 920) yhteisiä ovat (kanta-) verbin momentaanis-deminutiivinen verbijohdos *пāгьсэ* samoin kuin adjektiivi *пāгеч* (< ksa. **pākēhčē*). Adjektiiviin *поагкяй* pätee taas sama kuin adjektiiviin *суһтай*.

Sensiivijohdoksen *māӣйшэ* ’pitää sekopäisenä’ (vrt. *māяш* ’sekopäinen’, *māӣяй* id.) kantasana on SRS:n mukaan verbi *māӣе* ’olla sekopäinen’ (vastineita laajalti saamelaiskielissä; YSS § 1218).

Sanueeseen kuuluvista adjektiiveista *māйяй* on yksiselitteisesti preesenspartisiippi, kun taas *māяш* on monitulkintainen. Sana esiintyy myös KKLS:ssa, ja kummankin lähteen mukaan se on astevaihtelullinen (monikossa *māйяш*) eli taipuu kuten adjektiivi *вāйваш* (mon. *вāййваш*) ’köyhä’. Sanan analogisesti omaksumaa astevaihtelua on voinut edeltää kumpi tahansa johdostyypeistä, heikkoasteinen **tājōhčē* tai vahva-asteinen **tājōjēhčē*, mutta mahdollista lienee sekin, että se on syntynyt kummastakin seka-paradigmaisesti. Sensiivijohdoksen *цыԯԯкшэ* ’pitää kovaäänisenä’ välitön kantasana on adjektiivi *цыԯԯкай* ’kovaääninen’, joka ilmeisesti on preesenspartisiippi verbistä *цыԯԯкэ* ’helistä’.

Sensiivijohdos *сōрршэ* ’pitää sekaisin olevana’ kuuluu sanueeseen, jonka kantasana on verbi *сōррэ* ’sekoittaa’. SRS ei tunne tähän sanueeseen kuuluvia nomineja eikä sellaisia esitä myöskään KKLS. Verbin pohjoissaamelaisessa vastinesanueessa esiintyy myös substantiivi *sorri* ’sykkyrä, sotku (narussa)’ sekä adjektiivit *sorˈrái* ’sotkuinen, sotkeutuva (narusta)’ ja *soras* id. Kahdella ensin mainitulla on tarkka vastine inarinsaamessa, ja siellä esiintyy lisäksi substantiivi *sores* ’sotku (narussa)’. Adjektiivi *sorˈrái* on substantiivin *sorri* possessiivijohdos; adjektiivin *soras* suhteesta adjektiiviin *sorˈrái* ja verbiin *sorrat* ks. Nielsen 1979 [1926]: 215 (alav.), 224.

Verbikantainen on SRS:n ja KKLS:n valossa myös sensiiviverbi *рыһчшэ* ’pitää liian avonaisena (esim. vaatetta)’, vrt. *рыһче* ’avata (esim. verhot)’. Sanue on äänteellisen ja semanttisen variaationsa vuoksi historiallisesti monitulkintainen (vrt. YSS § 1038; Kuokkala 2018: 30), mutta sen yhteyteen kuulunee joka tapauksessa myös substantiivi In *ričče* ’alaston lapsi’. On mahdollista, että substantiivin alkuperäisempi merkitys on ollut ’repsottava vaate’ tjs., jolloin sanojen **rihčē* (subst.), ***rihčājȩ* (adj.) ja **rihčȩ-* (v.) morfologinen suhde on ollut sama kuin edellä puheena olleiden sanojen *sorri*, *sorˈrái* ja *sorrat*.

Sensiivijohdokset *вагкшэ* ’kokea kauhua, pelätä’, (vrt. *вагкай* ’kauhea’, *вагкшэсь* id.), *мӯдтшэ* ’pitää pilaantuneena’ (vrt. *мӯдтай* ’pilaantunut’) ja *лэ̄ншэ* ’pitää kovin nuhaisena’ (vrt. *лэ̄ннай* ’nuhainen’) kuuluvat sanueisiin, joille emme ole löytäneet vastineita SRS:n ulkopuolelta. Niiden kantasanoiksi SRS antaa substantiivit *вагк* ’kauhu’, *мӯдт* ’pilaantuneisuus’ ja *лэ̄нн* ’nuha’. Sanat *мӯдтай* ja *лэ̄ннай* ovat tämän aineiston valossa yksiselitteisesti possessiiviadjektiiveja. Samaa voi sanoa hieman varauksellisemmin myös sanasta *вагкай*. Sen rinnalla esiintyvä *вагкшэсь*, joka lienee attribuuttimuoto adjektiivista **вагкэш*, voi viitata siihen, että sanueeseen on kuulunut myös verbi, jolloin sana *вагкай* on johtosuhteiltaan samalla tavoin monitulkintainen kuin ylempänä käsitellyt *куарркэш* ja *суһтай*.

Sensiivijohdos *ōххтшэ* ’pitää jotakuta yksinäisenä’ kuuluu sanueeseen, jonka kantasana on numeraali *э̄ххт* ’yksi’. SRS ei tunne tähän sanueeseen kuuluvaa adjektiivia, jonka merkitys olisi ’yksinäinen’, mutta kylläkin adverbin *ōххтэ* ’yksin’, jonka tarkkoja vastineita ovat Ko *õhttu*, Po *okto* ja Et *aktegh* (Sammallahti 1998: 258). Tähän sanueeseen kuuluvaa tämänmerkityksistä adjektiivia ei tunne myöskään KKLS. Muista sanakirjoista löytyvät mm. Ko *õhttnaž*, In *oovthâs*, Po *oktonas*, *ovttaskas* ja Et *aktegs* ’yksinäinen’. On epäselvää, perustuuko kildininsaamen sensiivijohdos johonkin tämäntyyppiseen adjektiiviin, edellä mainittuun adverbiin vai suoraan numeraalikantasanaan.

Sensiivijohdoksen *руэпптшэ* ’pitää jotakuta toisin elävänä’ kantasanaksi SRS antaa adverbina ja prepositiona käytetyn sanan *руэппт* ’vastoin, vastaan’. SRS antaa sanan käytöstä mm. esimerkin *руэппт ēллей* ’toisella tavalla elävät’, jossa ensimmäinen sana tulee lähelle adjektiiviattribuuttia tai samanlaista yhdyssanan alkuosaa kuin on sanan vastineella inarinsaamen yhdyssanassa *ruáptukuánnil* ’vastavirta’. Samaan yhteyteen kuuluu pohjoissaamen *ruoktu* ’koti(paikka)’, jonka taustalla on konneksio *ruovttoluotta* ’(samaa tietä) takaisin’ (Sammallahti 1998: 261). Samantapaista käyttöä on myös sensiivijohdosten *поаллтшэ* ’pitää vieressä olevana’ ja *рӕввншэ* ’pitää reunalla olevana’ substantiivisilla kantasanoilla *поаллт* ’vieri’ ja *рēввьн* ’reuna’: *поаллт пэртэнь* ’vierekkäisissä taloissa’, *рēввьн пэ̄ррт* ’reunimmainen talo, reunatalo’.

Kaikkiaan voidaan todeta, että siinä kun muissa käsittelemissämme saamelaiskielissä sensiivisiä verbijohdoksia on muodostettu lähes pelkästään läpinäkymättömistä ”perusadjektiiveista”, on kildininsaamen kolmitavuisista sensiiviverbeistä valtaosa sekä muoto- että merkitysperustein välittömässä johtosuhteessa substantiivikantaiseen **-jȩ* -possessiiviadjektiiviin. Myös monitulkintaisissa tapauksissa on usein mahdollista ajatella kantasanaksi possessiiviadjektiivi, muutamissa tapauksissa taas verbikantainen adjektiivijohdos. Huomattavaa on, että nämä adjektiivityypit ovat lähes kaikki vartaloltaan vaihtelemattomasti vahva-asteisia, millä on ilmeisesti ollut vaikutusta kildininsaamen sensiiviverbityypin morfologiaan laajemminkin, kuten tuonnempana esitämme.

4. Nelitavuisista sensiiviverbeistä

Kaikissa saamelaiskielissä etelä- ja turjansaamea lukuunottamatta esiintyy myös vartaloltaan (alun perin) nelitavuisia sensiivijohdoksia, joiden johdin sisältää *š*-aineksen. Kaikki aineistoomme kuuluvat nelitavuiset sensiivijohdokset on lueteltu liitteessä 4. Yleensä johdin on muotoa **-(C)ušę* ja saa eri kielissä hieman erilaisia sidekonsonantteja kaksitavuisiin vartaloihin liittyessään. Kantasanat ovat adjektiiveja tai merkitykseltään adjektiivimaisia substantiiveja.

Luulajansaamessa sensiivinen johdin esiintyy muodossa *-lussjat* (Kintel 1991: 30, 54–55).[10] O. Korhosen (2005), Kintelin (2012) ja Grundströmin (1946–1954) sanakirjoissa on 23 kpl sensiivisiä *-lussjat*-johdoksia, joiden kantasanat ovat adjektiiveja (paitsi rajatapaukset *skádalussjat* 'harmitella' ← *skádá* 'vahinko' ja *suttalussjat* 'sääliä, surkutella' ← *suddo* 'synti'). Luulajansaamen johdoksista 13:lla (*áralussjat*, *binnalussjat*, *divralussjat*, *enalussjat*, *gávkalussjat*, *guhkalussjat*, *gåntsalussjat*, *nevralussjat*, *stuoralussjat*, *uhtsalussjat*, *unnalussjat*, *vájvalussjat*, *vastalussjat*)[11] on samassa kielessä myös kolmitavuinen rinnakkaismuoto ja kolmella (*nuoralussjat* 'nuoreksua', *amálussjat* 'vierastaa', *imálussjat* 'ihmetellä') on kolmitavuinen vastine pohjoissaamessa.

Ruong (1943: 170) mainitsee piitimensaamesta 13 sensiivistä *-luššat*-johdosta, joista 10:lla on samassa kielessä kolmitavuinen rinnakkaismuoto. Lisäksi Halászilta löytyy verbi *áralussjat* 'pitää aikaisena'. Piitimensaamen johdoksista 9 (muut paitsi *baskalussjat* 'pitää ahtaana', *giedtsalussjat* 'pitää kapeana', *hejulussjat* 'pitää köyhänä', *ånålussjat* 'pitää lyhyenä' ja *vánalussjat* 'pieneksyä') ovat samoja kuin luulajansaamessa. Uumajansaamen sanakirjoista löytyy vain yksi nelitavuinen sensiiviverbi, ja se vastaa tyypiltään piitimen- ja luulajansaamen johdoksia: *duvralussjat* 'pitää kalliina' ← *duvras* 'kallis' (Barruk 2018). Eteläsaamessa nelitavuiset sensiivijohdokset näyttävät olevan tuntemattomia.

Lähteestä KKLS on löytynyt 5 kildininsaamelaista nelitavuista sensiiviverbiä (asu tässä koltansaameen transponoituina): *ǩeäpsmõõššâd* 'pitää kevyenä', *puärsõõššâd* 'vanheksua', *jiâksõõššâd* 'oudoksua', *veârsõõššâd* 'vierastaa' ja *teevtõõššâd* 'ihmetellä'; näistä viimeisenä mainittu esiintyy myös lähteessä SRS (*мыфтуашэ*)[12]. Koltansaamen uudesta sanakirjasta (Moshnikoff & Moshnikoff 2021) löytyy neljä sensiiviseksi laskettavaa *-õõššâd*-verbiä: *occnjõõššâd* 'vähätellä' ← *occnjaž* 'vähäinen', *â'ǩkõõššâd* 'ikävöidä' ← *â'ǩked* 'ikävä', *õõmtõõššâd* 'ihmetellä' ← *õõmâs* 'ihme, kumma', *jâkstõõššâd* 'oudoksua' ← *jaakkâs* 'outo, vieras' (vrt. *jiâkstõõttâd* 'vierastaa'). Myös kildinin- ja koltansaamen johtimen asu on kantasaamen myöhäisvaiheessa ollut **-(C)ušę*. Nykypohjoissaamen aineistoissa (Nielsen, SSS) sensiiviset *-uššat*-verbit ovat harvinaisia; näihin voidaan lukea vain *aseǧuššat* 'pitää (liian) ohuena' (← *aseǧaš* 'ohut') ja varauksin *ahkiduššat* 'ikävöidä; pitkästyä' (← *ahkit* 'ikävä, pitkäveteinen').[13]

Inarinsaamesta (ILW) on *-šuššâđ*-päätteisiä sensiiviverbejä kirjattu kuusi kappaletta. Useimpia vastaa samanmerkityksinen kolmitavuinen *-šid*-verbi, joten verbityyppiä voi pitää kaksinkertaisena sensiivijohdoksena (esim. *asašuššâđ* 'pitää liian paksuna' ← *asašiđ* id. ← *assaa* 'paksu'). Verbin *puolâšuššâđ* 'pitää säätä liian kylmänä' rajakonsonantti voi osaltaan periytyä myös kantasanasta *puolâš* 'pakkanen'. Näiden lisäksi inarinsaamessa esiintyy *-šukšâđ*-päätteisiä sensiiviverbejä kuten *sevŋâdšukšâđ* 'pitää liian pimeänä' (*sevŋâd* 'pimeä'), ILW:ssa yhteensä 14 kpl. Toisin kuin *-uššâ*-aineksen sisältävät johdokset, -

[10] Kintel luokittelee johtimen myös essiiviseksi, nähtävästi lähinnä verbin *gávkalussjat* 'kjede seg' < *gávkas* 'kjedelig, trist' takia, joka johtosuhteeltaan kuitenkin kuuluu pikemmin sensiivisiin johdoksiin.

[11] Merkitykset: *áralussjat* 'pitää aikaisena', *binnalussjat* 'pitää pienenä', *divralussjat* 'pitää kalliina', *enalussjat* 'paljoksua', *gávkalussjat* 'pitkästyä' (← *gávkas* 'pitkäveteinen'), *guhkalussjat* 'pitää pitkänä, pitkästyä', *gåntsalussjat* 'pitää kummallisena', *nevralussjat* 'väheksyä, pitää huonona', *stuoralussjat* 'pitää liian isona', *uhtsalussjat* 'pieneksyä, väheksyä', *unnalussjat* 'pieneksyä', *vájvalussjat* 'pitää vaivalloisena', *vastalussjat* 'pitää rumana, inhota'.

[12] KKLS:ssa (s. 585) on lisäksi Genetziltä peräisin oleva akkalansaamelainen vastine *†divjuššə-*, jossa johtimen edellä oleva konsonantti on *j*.

[13] Muutoin *-(l)uššat*-johdokset ovat etupäässä essiivisiä, ilmaisevat kantasanan ominaisuudessa toimimista tai olemista (*gávváluššat* 'juonia, kieroilla', *guoktiluššat* 'kieroilla, vilpistellä', *skihpáruššat* 'kaveerata, olla kaveri(a)'), kuten useissa muissakin saamelaiskielissä.

šukšâđ-verbit ovat semantiikaltaan yksinomaan sensiivisiä. Myös johtimen liittyminen on erilaista (tästä tarkemmin luvussa 5).[14]

Muiden saamelaiskielten 1900-luvun sanastoissa ei *-kš*-aineksisia johdoksia esiinny lukuunottamatta Sammallahden (2021) uutta laajaa pohjoissaamen sanakirjaa, jossa esitetään suffiksihakusanan *-šakšit* 'moittia (ankarasti), haukkua' kohdalla itämurteesta kirjatut verbit *fuotninšakšit* 'haukkua huonoksi' (← *fuotni* 'huono'), *herskossakšit* 'haukkua liian herkuttelevaksi' (vrt. *herskostallat* 'herkutella', *herskui* 'herkutteleva'), *garassakšit* 'haukkua liian kovaksi' (← *garas* 'kova') ja *ruoinnassakšit* 'haukkua laihaksi' (← *ruoinnas* 'laiha').[15] Nämä verbit kuuluvat selvästi sensiivijohdoksiin, vaikkakin merkityksenä on negatiivista arviointia spesifimmin tällaisen arvion sanallinen ilmaiseminen.

1900-lukua vanhemmissa sanakirjoissa nelitavuisia sensiivijohdoksia esiintyy laajemmin. Friisin (1887) sanakirjasta löytyy 13 *-šokšat-*, 6 *-šavšat-* ja 7 *-šavšet*-loppuista sensiiviverbiä (esim. *akkedšokšat* : *-šovšam* 'anse for kjedelig' ← *akked* 'kjedelig, langsom'; *asašavšat* 'anse for tyk, for altfor tyk' ← *ăssai* 'tyk'; *divrašavšet* 'anse for dyr, for altfor dyr' ← *divres* 'dyr, kostbar'). Lähinnä itäruijalaisiin pohjoissaamen murteisiin pohjautuvassa Leemin (1768) sanakirjassa ei ole *-kš*-johdoksia, mutta siihen sisältyy kuusi nykypohjoissaamelle tuntematonta tyyppiä edustavaa *-šašša*-sensiivijohdosta, joista yhtä vaille kaikilla on rinnalla lyhyempi *-š*-johdos, esimerkiksi *amashjam* = *amashjashjam* 'anseer, holder for at være fremmed' (← *amas* 'fremmed'); *loijashjam* = *loijashjashjam* 'jeg holder for at være formeget stille og sagtmodig' (← *loigje* 'spag, sagtmodig, stille'). Lindahlin ja Öhrlingin (1780) lähinnä uumajansaamea edustavassa ruotsinsaamen sanakirjassa puolestaan esiintyy kolme *-haksjet*-loppuista verbiä, joista yksi on kiistatta sensiivinen (*wuorahaksjet* 'tycka eller hålla för gammal' ← *wuoras*, *wuores* 'åldrig, gammal'). Muiden kahden semantiikka on pikemminkin deminutiivista tai propinkvatiivista (*muitahaksjet* 'draga sig något litet till minnes'; *tåbdahaksjet* 'bära känsla till något, som man har förut sett, tycka sig känna igen något'). Myös Zacharias Plantinuksen 1670-luvulla laatimasta latinalais-eteläsaamelaisesta sanaluettelosta (ks. Koponen 2014) löytyy yksi nelitavuinen sensiivijohdos (*vastaiaxeth* 'detestari', vrt. LÖ *waste* 'stygg, ful'), mikä osoittaa, että tyyppi on aiemmin esiintynyt eteläsaamessakin.

5. Sensiivijohdosten derivotaksia

Pohjoissaamessa kolmitavuisten verbien sensiivijohdin *-š-* liittyy heikkoasteiseen kantavartaloon, jonka lopussa oleva vokaali on kantasanan vartalotyypistä riippuen (**e̯* >) *a*, (**ā* >) *á* tai (**u* >) *o*. Eteläsaamessa johtimen asu on ‹*sj*› ja sen edellä oleva vokaali vastaa historiallisesti pohjoissaamea. Koska eteläsaamesta puuttuu astevaihtelu, vartalomorfeemin aste ei ole määritelty. Piitimen- ja luulajansaamen sensiivijohdokset eroavat pohjoissaamesta siten, että johtimen asu on *-h-*, jonka edellä on (kantasaamen **e̯*:n jatkaja) *a*, myös silloin kun pohjoissaamessa on *á* (poikkeuksena vain Grundströmillä esiintyvä *stuoráhit*). Eri kielten johdosvastineiden tyypillisiä muotovastaavuuksia voidaan havainnollistaa 'pitää (liian) pitkänä' -verbillä (kantasana **kuhkē*), joka on myös ainut, joka on kirjattu kaikista aineistomme saamelaiskielistä:

Et	U	Pi	Lu	Po	In	Ko	Ki
gåhkasjidh	*guhkkáhit*	*guhkahit*	*guhkahit*	*guhkášit*	*kuhášiđ*	*kokkšed*	*куһкшэ*

Inarinsaamen kolmitavuiset sensiivijohdokset vastaavat derivotaksiltaan pohjoissaamea. Näin näyttää asia olevan KKLS:n suppean aineiston valossa koltan- ja kildininsaamessakin. Inarinsaamen sensiivijohdosta *rođášiđ* 'pitää rumana' vastaavassa kildininsaamen sanassa olisi johtimen edellä KKLS:n perusteella [*r⌒ð*$^{e}_{\breve{}}$*šeð*] kuitenkin (ollut) vokaaliin **ē* palautuva vokaali. SRS:ssa sana on asussa Ki *poaðuэ*, mikä vastaa sekä vartalomorfin heikon asteen että johtimen edellä olleen **ā*-vokaalin suhteen tarkalleen edellä mainittua inarinsaamen johdosta.

[14] Uudemmissa inarinsaamen sanakirjoissa (Sammallahti & Morottaja 1993; Olthuis et al. 2015–2022) ei nelitavuisia sensiiviverbejä esiinny. Lähelle sensiivistä semantiikkaa tulee *ahevuššâđ* ~ *ahevâššâđ* 'ikävöidä, kaivata', jonka objekti kuitenkaan ei ole "se, mitä pidetään ikävänä".

[15] Pekka Sammallahden mukaan tiedot ovat peräisin utsjokelaisilta vanhuksilta (sähköpostitiedonanto 15.4.2022).

Pohjoissaamen kanssa yhteisistä SRS:n johdoksista Ki *казшэ* ’pitää paksuna’, *нӯршэ* ’pitää nuorena’, *ражшэ* ’pitää heikkona’, *удцшэ* ’pitää pienenä’ ja *цуагшэ* ’pitää matalana’ vastaavat tarkalleen pohjoissaamen johdoksia *gasášit*, *nuorašit*, *rašášit*, *uhcášit*, *coagášit*, kun taas johdokset *вӣллкшэ* ’pitää vaaleana’, *ӣнншэ* ’paljoksua’, *кӭhпшэ* ’pitää kevyenä’, *коаммтшэ* ’pitää leveänä’, *куhкшэ* ’pitää pitkänä’, *кэ̄рршэ* ’pitää karkeana’, *лӣшшкшэ* ’pitää laiskana’, *нюццкшэ* ’pitää kosteana’, *о̄дтшэ* ’pitää uutena’, *поаhкшэ* ’pitää kuumana’, *пӯӣйтшэ* ’pitää rasvaisena’, *сӭӈукшэ* ’pitää ohuena’, *тӣллвшэ* ’pitää (pitkä)talvisena’, *туллвшэ* ’pitää runsasvetisenä’, *чӯррпшэ* ’pitää taitamattomana’ ja *ыййшэ* ’pitää liian öisenä’ ovat pohjoissaamelaisista vastineistaan poiketen vahva-asteisia (samoin *кӭнницшэ* ’pitää kapeana’, vrt. inarinsaamen *kiäzášiđ* id.). Vahva-asteisia ovat pääosin myös SRS:ssa esiintyvät muissa saamelaiskielissä tuntemattomat sensiivijohdokset. Odotuksenvastaiselle vartaloedustukselle lienee selityksenä se, että kun kildininsaamessa monilukuiset substantiivikantoja vastaavat johdokset ovat vaihtelemattomasti vahva-asteisen possessiiviadjektiivin pohjalta muodostettuja (esim. *пӯӈук* : gen. *пӯӈк* ’tuuli’ → *пӯӈькай* ’tuulinen’ → *пӯӈукшэ* ’pitää tuulisena’), on muissa sensiivijohdoksissa alkanut näiden mallivaikutuksesta myös ilmetä pyrkimys vahva-asteisuuteen.

Kolmitavuisten sensiivijohdosten kantasanat ovat pääosin adjektiiveja, joiden predikatiivimuoto on kaksitavuinen ja vokaaliloppuinen, mutta joukossa on muunkinlaisia. Niistä osalla on kaksitavuinen vokaaliloppuinen attribuuttimuoto (Po *ođđa* ’uusi’, *garra* ’kova’, *lossa* ’raskas’; vrt. pred.-muoto *ođas*, *garas*, *lossat*), mutta osalla attribuuttimuotokaan ei ole tällainen (Po *boaris*, attr. *boares* ’vanha’ ym.). Näissä tapauksissa johdos on siis muodostettu vokaaliloppuiseksi kaksitavuksi typistyneestä kanta-vartalomorfista (*boará-šit*). SRS:ssa on myös muutama ‹*сшэ*›-loppuinen sensiivijohdos, joiden kantasana on kaksitavuinen, sekä predikatiivi- että attribuuttimuodossa ‹*с*›-loppuinen adjektiivi: *пуаррсшэ*, *роаввсшэ*, *моаййвсшэ*, *моаввсшэ*, *вэрсшэ* (← *пуэресь* ’vanha’, *роавас* ’vahva’, *моайвас* ’taaja’, *моавас* ’vahva’, *вэрс* ’tuore’). On mahdollista, että kyseessä on hyperkorrekti ortografinen konventio pro **пуаршэ* (tai **пуаршэ*; vrt. Po *boarášit*) jne.[16] Mahdollista on toisaalta sekin, että nämä edustavat samaa tyyppiä kuin Friis *guorosset* (← *guoros* ’tyhjä’); äänneyhtymän *-sš-* osalta vrt. Friis *guorossokšat* (← *guoros* ’tyhjä’), Po *garassakšit* (← *garas* ’kova’), In *arvâssukšâđ* (← *aarvâs* ’antelias’).

Vokaaliloppuiseksi kaksitavuksi typistyneen kantavartalomorfin sisältävien kolmitavuisten sensiivijohdosten ohella pohjois-, inarin-, koltan- ja kildininsaamessa ja saamelaiskielten vanhoissa sanakirjoissa esiintyy nelitavuisia sensiivijohdoksia, joiden kantavartalomorfi sisältää kantasana-adjektiivin toisen tavun jälkeisen konsonantin. Tällaisia ovat yhtäältä Po *ahkidušša̮t* ’ikävöidä, pitkästyä’ koltansaamelaisine vastineineen (missä myöhäinen *d*-supistuma), Ki *puärsõõššad* ’pitää vanhana’ ja toisaalta In *viššâlšukšâđ* ’pitää ahkerana’, Friis 1887 *akkedšokšat* ’pitää ikävänä’. Ensin mainituissa sensiivijohdin on tyyppiä **-(C)ušę*, jälkimmäisessä tyyppiä **-(C)šukšę*. Inarinsaamessa ja vanhoissa sanakirjoissa esiintyvät näiden lisäksi tyypit **-(V)šušę* ~ **-(V)šęšę* (In *njuoskâšuššâđ* ’pitää raakana’, *poskâšuššâđ* ’pitää ahtaana’, Leem *buorashjashjam* ’pitää liian hyvänä (jollekin)’) ja **-(V)šukšę* ~ **-(V)šękšę* (In *sálttáášukšâđ* ’pitää suolaisena’, Friis *gukkašokšat* ’pitää pitkänä’, *asašavšat* ’pitää paksuna’).

Friisillä tavataan myös 4. tavultaan **ē*-vokaaliset varianttityypit **-(V)šękšē* (*divrašavšet* ’pitää kalliina’) ja **-(C)šękšē* (*apparšavšet* ’pitää isona’).[17] Jälkimmäinen näyttää esiintyvän yhä pohjoissaamen Utsjoen murteessa (*garassakšit*), missä konsonanttiloppuisen kantavartalo on ainakin yhdessä tapauksessa saatu kaksitavuisen kantasanan essiivimuodosta (*fuotninšakšit* ’haukkua huonoksi’ ← *fuotni* : ess. *fuotnin* ’huono’). Pekka Sammallahden sähköpostitiedonannon mukaan verbisuffiksi *-šakšit* toimii näissä nykykielen kannalta oikeastaan enkliittisenä sanana.

[16] Sama koskee sensiivijohdosta *мэӈусшэ*, jonka välittömäksi kantasanaksi olettaisi possessiiviadjektiivia (**мэӈусай* tjs. ← *мэӈс* ’varpu’ = Po *daŋas*).

[17] Lindahlin & Öhrlingin ja Plantinuksen eteläisen saamen sanastoissa nelitavuisten verbien vartalovokaalin alkuperäistä laatua ei pysty erottamaan (vokaalimerkkinä on aina ‹e›), mutta sensiivijohdosten tapauksessa on todennäköistä, että ne edustavat muualla laajalevikkistä **ę*-vokaalista tyyppiä.

6. Johdostyypin historiasta

Puheena olevan saamen sensiivijohtimen (esi)historiasta on kirjallisuudessa esitetty erilaisia käsityksiä. Sammallahti (1998: 93) pitää johdinta suomen samanfunktioisen suffiksin (esim. sanoissa *kummeksua*, *halveksia*) kielihistoriallisena vastineena rekonstruoiden sen esisaamelaiseksi asuksi *-*kśi* ja kantasaamelaiseksi *-*hčę*, kun taas E. Itkonen (1980: 27–28) katsoo saamen suffiksin olevan itämerensuomalaista lainaa. Jälkimmäiseen käsitykseen yhtyen (ks. myös Koponen 2013: 136) rekonstruoimme johtimelle kantasaamelaisen asun *-*kšę*, jonka konsonantismi heijastuu selvimmin inarinsaamen ja vanhojen sanakirjojen nelitavuisissa johdoksissa. Lainaoletusta tukee se, että kantasaamen *-*kš*-yhtymälle ei ole voitu osoittaa esisaamelaista taustaa, kun taas toisaalta itämerensuomen *s* on varsinkin *i*:n (ja muunkin suppean vokaalin) yhteydessä yleisesti lainattu saameen *š*:nä (esim. ksm. **silta* → ksa. * *šę̯ltē*, sm. *valmistaa* → saPo *válmmaštit*, sm. *matkustaa* → saPo *mátkkoštit*), ja siten *-*kšę* on suomen *-ksi*-johtimen (vrt. *halveksia*) odotettava lainavastine saamessa.

Nykyisissä saamelaiskielissä ja vanhoissa sanakirjoissa esiintyvien sensiivijohdostyyppien historialliset suhteet voi hahmotella seuraavanlaisesti (merkintä ~ tarkoittaa uuden variantin syntymistä edellisten rinnalle):

	I	II	III	IV	V	VI
vę̯stē →	vę̯stākšę		> vę̯stāšę	~ vę̯stāšukšę	~ vę̯stāšušę	
kę̯rę →	kę̯rę̯kšę		> kę̯rę̯šę	~ kę̯rę̯šukšę	~ kę̯rę̯šušę	
poarēs →	poarāsukšę	~ poarākšę	> poarāšę	~ poarāšukšę	~ poarāšušę	~ poarāsušę
ę̯kētē →	ę̯kētukšę			~ ę̯kētšukšę		~ ę̯kētušę
višę̯lē →	višę̯lukšę			~ višę̯lšukšę		~ višę̯lušę

Vaiheessa I johdin *-kšę* liittyy kaksitavuisiin adjektiivivartaloihin suoraan ja kolmitavuisiin adjektiivivartaloihin vartalovokaalin korvaavan sidevokaalin *u* (? ~ *ę̯*) välityksellä.[18] Jos adjektiivilla on sekä kaksitavuinen vokaaliloppuinen attribuuttimuoto (*kę̯rę̯*) että (pseudo)johtimen sisältävä vartaloltaan kolmitavuinen predikaattimuoto (*kę̯rę̯s* : *kę̯rę̯sę̯*), sensiivijohdin liittyy attribuuttimuotoon. Vaiheessa II syntyy tyyppi, jossa johtimen edellä on kantasanan kaksitavuinen vokaaliloppuinen allomorfi silloinkin, kun adjektiivilla ei ole tällaista attribuuttimuotoa. Vaiheessa III johdin lyhenee toisen ja kolmannen tavun rajalla asuun *-šę̯*. Vaiheessa IV syntyy pleonastinen[19] johdinvariantti *-šukšę̯*, jonka muodostimina ovat johtimen lyhentynyt ja lyhentymätön variantti. Vaiheessa V syntyy toinen pleonastinen johdinvariantti, joka sisältää kaksi lyhentynyttä varianttia ja vaiheessa VI syntyy tyyppi, missä johtimen lyhentynyt variantti on korvannut kolmannen ja neljännen tavun rajalla olleen pitkän johdinvariantin (**ę̯kētukšę̯* > **ę̯kētušę̯* jne.).[20]

Nykypohjoissaamessa sensiivijohdoksen valtatyyppi on (kaikkien saamelaiskielten tapaan) **vę̯stāšę̯-*, **kę̯rę̯šę̯-*, **poarāšę̯-*, minkä ohella siinä (samoin kuin koltansaamessa) esiintyy tyyppiä **ę̯kētušę̯-* edustava *ahkiduššat* (sekä *asehuššat*, jonka kanta-allomorfi on typistynyt karitiivijohtimen sisältävästä adjektiivista *aseheapmi*, *asehaš* 'ohut'). Tyyppiä **ę̯kētšukšę̯-*, **višę̯lšukšę̯-* edustavat In *muččâdšukšâđ* 'pitää kauniina', *viššâlšukšâđ* 'pitää ahkerana' ja Friis *akkedšokšat* 'pitää ikävänä', *gāvvelšavšat* 'pitää viekkaana'. Samaa **-Cšukšę̯*-tyyppiä ovat Friis *angeršavšat* 'pitää innokkaana',

[18] Jätämme esityksen selkeyttämiseksi tässä yhteydessä huomiotta sen, että osa nelitavuisista sensiivijohdoksista viittaa *u*:n sijasta *ę̯*-sidevokaaliin. Tarkoituksemme on käsitellä vokaalikysymystä tarkemmin toisaalla muiden (essiivisten ym.) **-(C)uše*-johdosten tarkastelun yhteydessä.

[19] Pleonasmista saamen historiallisen morfologian yhtenä kehitystendenssinä ks. Korhonen 1981: 274.

[20] E.Itkonen (1980: 28) pitää inarinsaamen **šušę̯*-tyyppiä joillakuilla puhujilla esiintyvänä myöhäisenä sekaantumana **-(C)ušę̯*-loppuisiin verbeihin, joiden johdin on muuta alkuperää (esim. *tuáhtáruššâđ* 'tohtoroida', *pevvârâššâđ* 'kokata'). Vaikka **-šušę̯*-tyyppi Leemin sanakirjan aineiston valossa (*loijashjashjam* ym.) näyttää Itkosen olettamaa vanhemmalta, varhaisemman sekaantumisen (tai muun vaikutuksen) mahdollisuutta ei voi sulkea pois. Palaamme tähänkin kysymykseen artikkelisarjamme kolmannessa osassa muiden **-(C)ušę̯*-johtimisten verbien käsittelyn yhteydessä.

guorossokšat ’pitää tyhjänä’ ja In *arvâssukšâđ* ’pitää antaliaana’. Friis *fastašavšet* ’pitää rumana, inhota’, *gărašavšat* ’pitää kovana’ ja *gæppašokšat* ’pitää kevyenä’ edustavat tyyppejä **vęstāšukšę-*, **kęrębukšę-*, **poarāšukšę-*, kun taas Leem *fastashjashjam* ’pidän rumana, inhoan’, *garashjashjam* ’pidän kovana’ ja In *kumášuššâđ* ’pitää liian kuumana’, *njuoskâšuššâđ* ’pitää raakana’ edustavat tyyppejä **vęstāšušę-*, **kęrębušę-* ja Ki *puärsõõššad* ’pitää vanhana’ edustaa tyyppiä **poarāsušę-*. Vanhoista ruotsinsaamen sanoista LÖ *wuorahaksjet* ’hålla för gammal’ edustaa tyyppiä **poarāšukšę-* ja Plant *vastaiaxeth* ’detestari’ (olettaen että siinä on kopiointivirhe: **vastahiaxeth*) edustaa tyyppiä **vęstāšukšę-*.[21] Luulajan-, piitimen- ja uumajansaamessa esiintyvän **-lušę-*johtimisen tyypin syntytapa jää epäselväksi. Sen *l* voisi olla tarttumaa jostakin **višęlušę-*tyyppisestä sanasta, mutta tällaisia ei näissä kielissä ainakaan nykyisin esiinny. Toisaalta mallina ovat saattaneet toimia eri johdostyyppiin kuuluvat **-(l)ušę-*verbit kuten Lu *gahppelussjat* ’auttaa synnytyksessä’ ← *gahppelit* id., *dåbdålussjat* ’olla tunnistavinaan’ ← *dåbddåt* ’tuntea’.[22]

Inarinsaamen sensiivijohdostyyppi *kimmáášukšâđ* (*kimmáá* ’kiimainen’), *sálttáášukšâđ* (*sálttáá* ’suolainen’) vastaa johtosuhteiltaan kildininsaamessa yleistä tyyppiä, jossa kantana niin ikään on possessiiviadjektiivi. Erona kildininsaameen on se, että inarinsaamen sanoissa on pleonastinen johdin, mikä puolestaan on aineistomme valossa kildininsaamelle tuntematon. Kuten kildininsaamen johdosten analyysin yhteydessä todettiin, osa kildininsaamen possessiiviadjektiivikantaisista sensiivijohdoksista on johtosuhteiltaan monitulkintaisia, koska niillä on tai voisi muiden saamelaiskielten aineiston valossa olla rinnalla merkitykseltään (lähes) identtisiä muuntyyppisiä adjektiiveja.

Pohjoissaamen Utsjoen murteessa ja Friisin sanakirjassa tavatut vokaalivarianttityypit Po *garasšakšit*, Friis *fastašavšet* vaikuttavat neljännen tavun vokaalinsa valossa vaihtaneen myöhäisesti taivutustyyppiä: alkuperäisemmän johdostyyypin (*)*fastašakšat* yksikön preesenstaivutus *fastašavššan* : *fastašavššat* : *fastašakšá* on siirtynyt äänteellisesti läheiseen **ē*-vartaloiseen tyyppiin *fastašavššán* : *fastašavššát* : *fastašakšá*. Toinen analogiakehitys näkyy johtimen heikkoasteisen *-všš-*konsonantismin yleistämisessä Friisin tyypeissä *fastašavšet*, *apparšavšet*.

Produktiivinen sensiivijohdostyyppi, jossa kantasanana on substantiivista johdettu possessiiviadjektiivi, on aineistomme valossa kildininsaamelainen (tai ehkä vain lähteen SRS) innovaatio, mutta sen juuret voivat ulottua kildininsaamen erilliskehitystä edeltävään aikaan. Tyypin lähin vastine on yllä mainittu inarinsaamen pleonastinen *sálttáášukšâđ*-tyyppi, mutta myös kildininsaamen ulkopuolella on sensiivijohdoksia, joiden kantasana voisi olla possessiiviadjektiivi. Tällainen on esimerkiksi pohjoissaamen *dulvvášit*, jonka kantasana voi olla johtamaton substantiivi *dulvi*, sen muuntyyppinen adjektiivijohdos *dulvvis* tai possessiiviadjektiivi *dulvái*, samoin kuin johtosuhteiltaan tähän rinnastuva pohjoissaamen *sálttášit*, vrt. *sálttis* ja *sáltái*.[23] Kildininsaamen aineistossa on merkkejä siitä, että possessiiviadjektiivikantaisia sensiiviverbejä olisi alettu hahmottaa ensisijaisesti johtoketjun taustalla olevan substantiivin johdoksiksi, ja uudelleenanalyysin jälkeen vastaavia johdoksia on alettu muodostaa myös suoraan substantiiveista. Nähtävästi vaihtelemattomasti vahva-asteisten possessiiviadjektiivien vaikutuksesta on johdostyyppi alkanut liukua myös vanhojen adjektiivikantaisten johdosten osalta vahva-asteista kantavartaloa edellyttäväksi.

7. Lopuksi

Tarkastelumme perusteella voidaan todeta, että eri saamelaiskielten *š-* (kielittäin > *h-*)aineksiset sensiiviset verbijohdostyypit ovat palautettavissa kantasaamelaiseen **-kšę-*johtimeen, joka puolestaan on

[21] Friis *guorosšet* ja SRS *nyappcшэ* jne. edustavat taulukosta puuttuvaa tyyppiä **poarāssę-*, joka (sikäli kuin ei ole sanakirjan laatijoiden teoretisoima) on syntynyt analogisesti mallin Friis *gukkašokšat* = *gukkašet* mukaan.

[22] Ruongin (1943: 170) mukaan ainakin jotkut piitimensaamelaiset informantit pitivät *-lušša*-sensiivijohdoksia vähäisempää määrää osoittavina kuin *-š-*johdoksia, ikään kuin *-l-* assosioituisi jollain lailla *-l-*subitiiveihin. Subitiivijohtimesta ei tässä kuitenkaan voi olla suoraan kyse, koska johdin on verbikantainen ja edellyttää vahva-asteista kantavartaloa.

[23] Tämän tutkimuksen ulkopuolelle jäävä kysymys on, olisiko (pohjois)saamen adjektiivityyppi *dulvvis*, *sálttis*, *suttis* (samoin kuin *dulppas*, *soras*) possessiiviadjektiivin edelleenkehittymä.

itämerensuomesta saatu lainasuffiksi. Kantasaamelaiseen taustaan viittaavat selvästi kymmenkunta etelä-, keski- ja itäryhmän kielille yhteistä johdosta. Erot eri kielissä esiintyvien johdosten määrissä selittyvät osaksi aineistopohjan epätasaisuudella, mutta on ilmeistä, että kolmitavuinen sensiivijohdostyyppi on kielten erilliskehityksen aikana muuttunut erittäin produktiiviseksi toisaalta pohjois- ja inarinsaamessa, toisaalta kildininsaamessa. Epävarmaksi jää, viittaavatko näille keski- ja itäryhmän kielille yhteiset parikymmentä johdosta jonkinlaiseen yhteiseen kehitysvaiheeseen (ja johdostyypin myöhempään harvinaistumiseen koltassa) vai onko kyse vain produktiivisuuden aiheuttamista yhteensattumista.

Nelitavuisissa johdoksissa, joissa saamen johdinten alkuperäisen muodon yleensä olettaisi painosuhteiden takia säilyvän parhaiten, kehitys on johtanut monenlaisiin pleonastisiin ja analogisiin muodostustyyppeihin, ja *kš*:llisestä johdinvariantista on säilynyt jälkiä vain inarinsaamessa, joissakin pohjoissaamen itämurteissa sekä vanhoissa kielenkuvauksissa. Kuvaavaa muutosten nopeudelle on, että vielä 1800-luvulla pohjoissaamen sanakirjoissa ja kieliopeissa esiintynyt *-šokšat ~ -šavšat ~ -šakšit* -johdostyyppi on muuttunut harvinaiseksi murrereliktiksi, eikä inarinsaamenkaan vastaavaa tyyppiä löydy enää uusimman kielen aineistoista.[24] Myös kildininsaamen kolmitavuisten johdosten muodostuksessa possessiiviadjektiivien vaikutuksesta syntynyt suuntaus vartalon vahva-asteisuuteen havainnollistaa, kuinka hyvinkin satunnaiset muodostustyyppien yleistymiset voivat johtaa laajempaan morfologian muutokseen.

Kielten lyhenteet

Et = eteläsaame	ksa. = kantasaame	Po = pohjoissaame
In = inarinsaame	ksm. = kantasuomi	T = turjansaame
Ki = kildininsaame	Lu = luulajansaame	U = uumajansaame
Ko = koltansaame	Pi = piitimensaame	

Aineistolähteet kielittäin

(Suluissa mahdollinen aineistotaulukoissa käytetty lyhenne. Ensimmäisenä mainittu kunkin kielen päälähde on aineistotaulukossa ilman merkintää.)

Eteläsaame: ÅaDB, LW
Uumajansaame: Schlachter 1958, Barruk 2018 (B)
Piitimensaame: Ruong 1943, Wilbur 2020 (W), Halász 1896 (H), LW
Luulajansaame: Korhonen 2005, Kintel 2012 (AK), Grundström 1946–1954 (G), LW
Pohjoissaame: SSS, Sammallahti 2021 (S2), Nielsen (N), LW
Inarinsaame: ILW
Koltansaame: Moshnikoff & Moshnikoff 2021, Sammallahti & Mosnikoff 1991, KKLS
Kildininsaame: SRS, KKLS
Vanhat sanakirjat ja -luettelot: Friis (1887), Leem (1768), LÖ (Lindahl & Öhrling 1780), Plant (Plantinus 1670-l.)

Lähteet

Barruk, Henrik. 2018. *Báhkuogirjjie: Ubmejesámien–dáruon, Dáruon–ubmejesámien / Ordbok: Umesamisk–svensk, Svensk–umesamisk*. Umeå.

[24] Myöskään SIKOR-korpuksen (https://gtweb.uit.no/korp/) pohjois- ja inarinsaamen nykykirjakielten aineistoista ei löydy tämän tyypin esiintymiä (haettu sananmuotoja säännöllisellä lausekkeella ”.+š[aeou]+[kv]+š.+” 15.4.2022).

Friis = Friis, J. A. 1887. *Lexicon lapponicum / Ordbog over det lappiske sprog*. Christiania.
Grundström, Harald. 1946–1954. *Lulelapsk ordbok / Lulelappisches Wörterbuch* 1–4. (På grundval av K. B. Wiklunds, Björn Collinders och egna uppteckningar utarbetad av Harald Grundström. Skrifter utgivna genom Landsmåls- och Folkminnesarkivet i Uppsala, Ser. C:1.) Uppsala.
Halász, Ignácz. 1896. *Pite lappmarki szótár és nyelvtan*. (Svéd-lapp nyelv VI.) Magyar Tudományos Akadémia, Budapest.
ILW = Itkonen, Erkki. 1986–1989. *Inarilappisches Wörterbuch* 1–3. (Lexica Societatis Fenno-Ugricae 20.) Suomalais-Ugrilainen Seura, Helsinki.
Itkonen, Erkki. 1980. Über einige lappische Verbalsuffixe. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 76: 23–41.
Kintel, Anders. 1991. *Syntaks og ordavledninger i lulesamisk*. Samisk utdanningsråd, [Kautokeino].
Kintel, Anders. 2012. *Julevsáme-dárro báhkogirjje*. http://gtweb.uit.no/webdict/ak/smj2nob/ (luettu 2020-11-25).
KKLS = Itkonen, T. I. 1958. *Koltan- ja kuolanlapin sanakirja / Wörterbuch des Kolta- und Kolalappischen* I–II. (Lexica Societatis Fenno-Ugricae 15.) Suomalais-Ugrilainen Seura, Helsinki.
Koponen, Eino. 2013. Beiträge zur uralischen Wortbildungslehre mit besonderer Berücksichtigung des Samischen. *Finnisch-Ugrische Mitteilungen* 37: 77–159.
Koponen, Eino. 2014. Zur (Vor-)Geschichte der saamischen Lexikografie: Ein lateinisch-saamisches Wörterverzeichnis aus dem 17. Jahrhundert. In *Proceedings of the XVI EURALEX International Congress: The user in focus*. 15–19 July 2014, s. 749–765. Bolzano/Bozen. https://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_057_p_749.pdf.
Koponen, Eino & Kuokkala, Juha. 2021. Kantasaamen *-*(ę)hčę*-frekventatiivijohtimen edustuksesta nykyisissä saamelaiskielissä. In *Multilingual Facilitation: Honoring the career of Jack Rueter*, s. 187–196. Toim. Mika Hämäläinen & Niko Partanen & Khalid Alnajjar. Helsinki. https://doi.org/10.31885/9789515150257.
Korhonen, Mikko. 1981. *Johdatus lapin kielen historiaan*. Suomalaisen Kirjallisuuden Seura, Helsinki.
Korhonen, Olavi. 2005. *Báhkogirjje: Julevusámes dárrui, dáros julevusábmái / Lulesamisk svensk, svensk lulesamisk ordbok*. Sámij åhpadusguovdásj, Jokkmokk.
Kuokkala, Juha. 2018. Finnic-Saamic labial vowels of non-initial syllables: An etymological evaluation. In *Περὶ ὀρθότητος ἐτύμων: Uusiutuva uralilainen etymologia*, s. 11–76. Toim. Sampsa Holopainen & Janne Saarikivi. (Uralica Helsingiensia 11.) Suomalais-Ugrilainen Seura, Helsinki.
Leem = Leem, Knud. 1768. *Lexicon Lapponicum bipartitum, Lapponico-Danico-Latinum & Danico-Latino-Lapponicum cum Indice latino. Pars Prima Lapponico-Danico-Latina*. Seminarium Lapponicum Fridericianum, Nidrosiae.
LW = Lagercrantz, Eliel. 1939. *Lappischer Wortschatz* I–II. (Lexica Societatis Fenno-Ugricae 6.) Suomalais-Ugrilainen Seura, Helsinki.
LÖ = Lindahl, Erik & Öhrling, Johannes. 1780. *Lexicon Lapponicum*, cum interpretatione vocabularum sveco-latina et indice svecano-lapponico. Holmiae. (Uudistettu digitaalinen versio: http://www.raamesuenne.se/Lexicon_lapponicum_20160330.pdf)
Moshnikoff, Satu & Moshnikoff, Jouni. 2021. *Suomi–koltansaame-sanakirja / Lää'dd-sää'm sää'nnǩe'rjj*. Toim. Miika Lehtinen, Eino Koponen, Merja Fofonoff ja Raija Lehtola. Sää′mte′ğğ, Aanar. (Verkkoversio käytettävissä osoitteessa https://saanih.oahpa.no/fin/sms/, aineisto haettu osoitteesta https://gtsvn.uit.no/langtech/trunk/words/dicts/finsms/src/ [revisio 193190, 2021-04-22])
Nickel, Klaus Peter. 1990. *Samisk grammatikk*. Samisk utdanningsråd, [Oslo].
Nickel, Klaus Peter & Sammallahti, Pekka. 2011. *Nordsamisk grammatik*. Davvi Girji, Karasjok.
Nielsen = Nielsen, Konrad. 1979 [1932–62]. *Lappisk (samisk) ordbok / Lapp dictionary* 1–5. Based on the Dialects of Polmak, Karasjok and Kautokeino. (Instituttet for sammenlignende kulturforskning, Serie B: 17.) Oslo.
Nielsen, Konrad. 1979 [1926]. *Lærebok i lappisk* I: Grammatikk. Universitetsforlaget, Oslo.

Olthuis, Marja-Liisa & Valtonen, Taarna & Seurujärvi, Miina & Trosterud, Trond. 2015–2022. *Nettidigisäänih Anarâškiela-suomakielâ-anarâškielâ sänikirje.* UiT, Tromsø. https://saanih.oahpa.no/smn/fin/ (Aineisto haettu osoitteesta https://gtsvn.uit.no/langtech/trunk/words/dicts/smnfin/src/ 2022-04-14)
Plant = Plantinus, Zacharias. (1670-luku?) (Latinalais-saamelainen sanaluettelo. Julkaistu artikkelissa Setälä 1890.)
Ruong, Israel. 1943. *Lappische Verbalableitung dargestellt auf Grundlage des Pitelappischen*. Uppsala.
Rydving, Håkan. 2013. *Words and varieties: Lexical variation in Saami*. (Suomalais-Ugrilaisen Seuran Toimituksia 269.) Suomalais-Ugrilainen Seura, Helsinki.
Sammallahti, Pekka. 2002. *North Saami resource dictionary*. (Publications of the Giellagas Institute 1.) Oulu.
Sammallahti, Pekka. 1998. *The Saami languages: An introduction*. Davvi Girji, Kárášjohka.
Sammallahti, Pekka. 2021. *Sámi – suoma sátnegirji / Pohjoissaame–suomi-sanakirja*. Verkkoversio: http://satni.org/sammallahtismefin (luettu 2022-03-28).
Sammallahti, Pekka & Morottaja, Matti. 1993. *Säämi-suomâ sänikirje / Inarinsaamelais-suomalainen sanakirja*. Girjegiisá, Ohcejohka.
Sammallahti, Pekka & Mosnikoff, Jouni. 1991. *Suomi-koltansaame sanakirja / Lää'dd-sää'm sää'nnǩe'rjj*. Girjegiisá, Ohcejohka.
Schlachter, Wolfgang. 1958. *Wörterbuch des Waldlappendialekts von Malå und Texte zur Ethnographie.* (Lexica Societatis Fenno-Ugricae XIV.) Suomalais-Ugrilainen Seura, Helsinki.
Setälä, E. N. 1890. Ein lappisches wörterverzeichnis von Zacharias Plantinus. *Suomalais-Ugrilaisen Seuran Aikakauskirja* 8: 85–104.
SRS = Куруч, Р. Д. (toim.). 1985. *Саамско-русский словарь / Сӑм-рӯшш соагкнэ̑нкь*. Русский язык, Москва. (Verkkoversio: http://slovari.saami.su/slovari/saamsko-russkij-slovar-kuruch/Saami-Russian-dictionary-Kuruch-1/)
SSS = Sammallahti, Pekka. 1989. *Sámi-suoma sátnegirji / Saamelais-suomalainen sanakirja*. Jorgaleaddji, Ohcejohka.
Wilbur, Joshua. (toim.). 2020. *Bidumsáme Báhkogirrje / Pitesamisk ordlista / Pite Saami Word List* [online-tietokanta]. http://saami.uni-freiburg.de/psdp/pite-lex/ (luettu 2020-04-24).
VISK = Hakulinen, Auli & Vilkuna, Maria & Korhonen, Riitta & Koivisto, Vesa & Heinonen, Tarja Riitta & Alho, Irja. 2004. *Iso suomen kielioppi*. Suomalaisen Kirjallisuuden Seura, Helsinki. Verkkoversio: http://scripta.kotus.fi/visk.
YSS = Lehtiranta, Juhani. 1989. *Yhteissaamelainen sanasto*. (Suomalais-Ugrilaisen Seuran Toimituksia 200.) Suomalais-Ugrilainen Seura, Helsinki.
ÅaDB = Bergsland, Knut & Magga, Lajla Mattsson. 1993. *Åarjelsaemien-daaroen baakoegærja / Sydsamisk-norsk ordbok*. Iđut, [Lakselv].

LIITE 1. USEAMMASSA KIELESSÄ ESIINTYVÄT 3-TAVUISET SENSIIVIJOHDOKSET

kantasana	merkitys	LW	Etelä	Uumaja	Piitime	Luulaja	Pohjois	Inari	Koltta	Kildin
1. Etelä-, keski- ja itäryhmässä tavatut johdokset										
čuorbi	taitamaton	784			tjuorbahit		čuorbbášit	čuárbášiđ		чӯррпшэ*
eatnat	paljon	866	jeanasjidh		ienahit	enahit G	eanášit	iänášiđ	jäänšed	ӓнншэ*
gassat	paksu	1987			gasahit		gasášit	kasašiđ		казшэ*
kezzi In	kapea	2394			giedtsahit			kiäzášiđ		кӓннцшэ
guhkki	pitkä	2776	gåhkasjidh	guhkkáhit	guhkahit	guhkahit	guhkášit	kuhášiđ	kokkšed	куњкшэ*
oavdu	ihme	4602			åvduhit H	oavdduhit	ōvddošit		ååudšed	
buorrę	hyvä	5323	buarasjidh				buorášit		puâršed	puâršed
uhcci	pieni	8228	åhtjasjidh		uhtsahit H	uhtsahit	uhcášit	ucášiđ	occšed	уддшэ*
vánis	niukka	8432			vánahit		vánášit	vánášiđ		väänšed
2. Etelä- ja keskiryhmässä tavatut johdokset										
headju	heikko	1320			hejuhit		hējošit	hiäjušiđ		
nuoges Lu	riittävä	4275	nuakasjidh	nuahkáhit		nuohkahit				
baski	ahdas	4691			baskahit		baskkášit			
binˈná	pala; vähäinen	4937			binnahit	binnahit				
stuoris	suuri	7074	stuarasjidh			stuoráhit G	stuorášit	stuárášiđ		
divrras	kallis	7872		dïvrahit		divrahit	divrrašit	tivrâšiđ		
unˈni	pieni	8252	ånnasjidh		unnahit	unnahit G	unnášit			
fasti	ruma	8364	vastasjidh		vastahit H	vastahit	fasttášit	vastašiđ		
3. Keski- ja itäryhmässä tavatut johdokset										
amas	outo	79					amašit	omâšiđ	õõmšed	õõmšed
coahki	matala (vesi)	401					coagášit			цуагшэ
hálbi	halpa	1213					hálbbášit	hálbášiđ	häälbšed	
idja	yö	1555					ijašit	ijâšiđ		ыййшэ
garas	kova	1969					garašit	korâšiđ		кӛрршэ
geahppat	kevyt	2272					geahpášit N	kiäpášiđ		кӓњпшэ
govdat	leveä	2499					govddášit	kobdášiđ		кӧммтшэ
láiki	laiska	3222					láikkášit	láškášiđ		лӓшшкшэ*
loikkas	haalea	3489					loikkašit	luškâšiđ	looškšed	looškšed
lossat	painava	3518					losášit		lååzzšed	lååzzšed
nuorra	nuori	4281					nuorašit	nuorâšiđ		нӯршэ*
njuoskkas	märkä	4471					njuoskkašit	njuoskâšiđ		нюӧццкшэ
ođas	uusi	4481					ođašit	uđâšiđ		ӧдтшэ*
oamis	vanha	4573					oamášit			vuämšed
báhkka	helle	4746					báhkašit			поањкшэ
boaris	vanha	5149					boarášit			пуаррсшэ
buoidi	lihava	5299					buoiddášit	puáidášiđ		пӯййтшэ
rašši	hauras	5502					rašášit	rašašiđ		ражшэ
seaggi	hoikka	6180					seakkášit	siägášiđ		сӓӈукшэ
dálvi	talvi	7728					dálvvášit N			тӑллвшэ
dulvvis	tulvainen	8026					dulvvášit			туллвшэ
vielgat	valkoinen	8674					vielggašit			вйллкшэ
vuogas	sopiva	8760					vuogašit			вӯгсшэ
	ruma	-						rođášiđ		роадшэ*
4. Vain keskiryhmässä tavatut johdokset										
asˈsái	paksu	208					asášit	asašiđ		
árgi	arka	228					árggášit	árgášiđ		
árrat	varhainen	241				árahit	árašit			
árvvas	antelias	248					árvvašit	arvâšiđ		
čáppat	kaunis	496				tjábbahit	čáppášit			
fuotni	huono	1121					fuonášit	huánášiđ		
hánis	saita	1234					hánášit	hánášiđ		
jalˈla	hölmö	1620					jallašit	jolâšiđ		
galmmas	kylmä	1903					galmmašit	kolmâšiđ		
gáhcci	kitsas	2018					gáhcášit N	kácášiđ		
geafi	köyhä	2238					geafášit	kiävhášiđ		
goikkis	kuiva	2517					goikkášit	koškášiđ		
guoirras	laiha	2873					guoirrašit	kuoirâšiđ		
nanus	luja	4070					nanošit	nanošiđ		
nealgi	nälkä	4143					nealggášit	niälgášiđ		
neavri	surkea	4157				nevrahit	neavrrášit N	niävrášiđ		
nuoski	siivoton	4286					nuoskkášit	nuáskášiđ		
njárbat	harva	4357					njárbbášit	njárbášiđ		
njuolgat	suora	4461					njuolggašit	njuolgâšiđ		
oppas	umpinainen	4478					oppašit	ubâšiđ		
bahča	katkera	4616					bahčašit	počâšiđ		
rákkis	kireä	5528					rákkášit	rágášiđ		
rávža	kitulias	5603					rávžžašit	ravžâšiđ		
rikkis	rikas	5707					rikkášit	rigášiđ		
ruoinnas	laiha	5941					ruoinnašit	ruoinâšiđ		
sálti	suola	6116				sáltahit G	sálttášit			
seavdnjat	pimeä	6224					seavnnjášit	siävnjášiđ		
silli	laiha	6263					silášit	silášiđ		
dušˈši	turha	8049					dušášit	tušášiđ		
udju	ujo	8235					ujošit	ujošiđ		
váivi	vaiva	8400				vájvahit	váivvášit	váivášiđ		
gávkkas	toljottava	-				gávkahit	gávkkašit N			
		73	8	3	12	15	69	54	9	31

LIITE 2. VAIN KILDININSAAMESSA ESIINTYVÄT 3-TAVUISET SENSIIVIJOHDOKSET

sens.-johdos	johd:n k.-s.	sanueen k.-s.	merkitys	KKLS
1. kantasanana possessiiviadjektiivi				
оаллшэ	а̄ллай	а̄лль	kosteus	10
оаӈӈкшэ	оаӈӈкай	оаӈӈк	hanki	14
аббршэ	аббрай	аббьр	sade	17
ыллшэ	ылляй	ылл	hiili	44
йӈӈкшэ	йӈӈҍкай	йӈӈк	henki	58
ёӈӈшэ	ёӈӈяй	ёӈӈ	puolukka	68
кэ̄бпшэ	кэ̄бпай	кэ̄бп	sairaus	107
кунншэ	куннай	кунн	tuhka	165
куэллшэ	куаллай	кӯлль	kala	172
куэрркшэ	куэрркай	куэррк	matalikko	175
лоастшэ	la'stti Ko	лоасст	lastu	195
лаппьсшэ	лаппсай	лаппьс	kaste	200
ля̄ййпшэ	ля̄ййпай	лёййп	leipä	204
лыййнншэ	лыййнай	лыййн	pellava	213
лоантшэ	låddai	лоаннҍт	lintu	218
луэдтшэ	луадтай	луэдт	vaahto	226
луэҥкшэ	луэҥкай	луэҥк	mäki	226
лӯзшэ	лӯссай	лӯсс	lohi	228
мырршэ	мырряй	мырр	sopu	256
моаҥкшэ	моаҥкай	моаҥкь	mutka	259
моарршэ	mårrai 902	моаррь	kura	260
мулӆтшэ	мулӆтай	мулӆьт	saippua	263
мӯнньшэ	мӯнняй	мӯнь	pakkanen	263
мӯрршэ	мӯрряй	мӯрр	puu	265
мӯррьйшэ	мӯррьяй	мӯррьй	marja	265
нёаллшэ	нёаллай	нёалл	naali	293
ниввлшэ	ниввьляй	ниввьл	lima	303
па̄нншэ	па̄ннэ̄й	па̄ннҍ	hammas	338
па̄ррншэ	пя̄ррнай	*па̄ррьн	lapsi	340
поарршэ		поарр	höyry	341
пя̄ллтшэ		пя̄ллт	aukea (s.)	348
пэ̄ввлшэ	пэ̄ввлай	пэ̄ввл	pilvi	353
пйӈӈкшэ	пйӈӈькай	пйӈӈк	tuuli	363
пйввлшэ	пйввлай	пйввл	pälvi	375
пуазшэ		пуаз	poro	382
по̄ннцшэ	по̄ннҍцай	по̄ннц	sulka	388
по̄рркшэ	по̄ррькай	по̄ррк	pyry	396
поаввншэ	поаввнай	поаввьн	mätäs	400
ра̄ййкшэ	ра̄ййкай	ра̄ййк	reikä	416
ра̄зшэ	ра̄ссяй	ра̄ссь	ruoho	422
руэммпшэ	ruâbbai	рӯммьп	rupi	455
руэсскшэ		руэсск	verkakoriste	457
рӯсстшэ	ruõ'stti	рӯсст	ruoste	458
са̄кксшэ	са̄кксай	са̄ккьс	lika	464
са̄рркшэ	са̄ррькай	са̄ррьк	haara	475
сйййшэ	sieddjái Po	сййй	visva	489
сӯкксшэ	сӯккьсай	сӯккс	toukka	527
сӯййншэ	сӯййнай	сӯййн	heinä	528
суэллшэ	суэллай	сӯлль	suola	530
суввшэ	суввяй	сувв	savu	536
субпшэ	субпай	субпь	haapa	537
та̄ххтшэ	та̄ххтай	та̄ххьт	luu	566
таррвшэ	таррвай	таррьв	terva	573
тэӈӈсшэ		тэӈс	varpu	582
тя̄ннкшэ	тя̄ннкай	тённҍк	raha	587
тя̄ррмшэ	тя̄ррмай	тёррьм	törmä	590
цыввншэ		цыввьн	virtsanhaju	635
цӯннцшэ	цӯннҍцай	цӯннц	kalvo	638
ча̄дзшэ	ча̄дзай	ча̄дзь	vesi	649
чйврршэ	чйврряй	чйврр	somero	669
чиввлшэ	čeullai	чиввьл	syylä	669
чӯшшкшэ	čuõ'škki	чӯшшк	sääski	693
уррмшэ	уррмай	уррьм	permu	701
вйгкшэ	вйгкяй	вйгк	voima	740

sens.-johdos	johd:n k.-s.	merkitys	sanueen k.-s.	merkitys	KKLS
2. Kantasanana muunlainen adjektiivi					
а̄ййвшэ	а̄ййвэсь	hiljainen			7
ёҥкшэ	ёҥк	vieras			48
кыннтшэ	кынтас	tiukka			114
кӯдцшэ	кӯдц	pilaantunut			179
мйццкшэ	мйццк	laho			254
мёвшэ	мёвв	pieni			256
моажьшэ*	мо̄джесь	kaunis			262
нёллкшэ	нёллькесь	tasainen			308
нюазшэ*	нюэссь	huono			310
э̄ллшэ	э̄лл	korkea			316
поассшэ	поасс	paha			342
пя̄ццкшэ	пя̄ццк	jyrkkä			373
пунншэ	пунншэсь	siisti			406
роаввсшэ	роавас	vahva			426
руэнншэ	руэнн	vihreä			457
руэпсшэ	рӯппьсесь	punainen			460
ручччкшэ	ручччкесь	keltainen			461
сыннҍшэ	сыннҍ	saita			485
суанншэ	суэнЭсь	rauhallinen			532
шйгкшэ	шйгк	hyvä			550
шӯрршэ*	шӯрр	iso			563
тоаййвсшэ	тоайвас	taaja			576
тоаввсшэ	тоавас	vahva			577
тйввтшэ	тйввт	täysi			595
чабшэ	чаб	ehyt			651
чёҥпьшэ	чёҥпь	taitava			663
чо̄гкшэ	чо̄гк	terävä			676
вэрсшэ	вэрс	tuore			735
вӯйкшэ	вӯййк	suora			778
вуэгкшэ	вӯгкесь	tuuhea			782
вуэллшэ	вӯллькесь	matala			786
нёммшэ	нёмм	veltto			
рёввкшэ	ря̄ввк	juhlava			
цыййшэ	цыйя	hyvä (lapsi)			
		3. Monitulkintaiset			
о̄ххтшэ			э̄ххт	yksi	29
яллшэ	яллай	iljanteinen	ялл	iljanne	48
ямкшэ	ёҥк	vieras			48
ка̄ййншэ	ка̄ййнай	ujo	ка̄йнэч	ujo	81
куарркшэ	куарркэш	kerskaileva	куаррк	kerskailu	135
мйллшэ	мйллэв	älykäs	мйлл	äly	252
поаллтшэ			поаллт	vieri	334
ря̄ввншэ			рёввьн	reuna	434
рыҥчшэ			рыҥче	avata	443
руэпптшэ			руэппт	vastaan	460
со̄рршэ			со̄ррэ	sekoittaa	516
суннтшэ			суннҍт	sulapaikka	525
суҥтшэ	суҥтай	oikukas	суҥт	oikku	537
та̄ййшэ			та̄йе	taajoa	568
то̄ӆӆкшэ	то̄ӆӆкэш	älykäs	то̄ӆӆк	äly	604
то̄ллшэ			то̄лл	tuli	605
тӯллпшэ	тӯлльпай	tylppä	тӯллп	tylppä	606
тоасскшэ	тоасскай	surullinen	тоасськ	suru	609
цыӆӆкшэ			цыӆӆкэ	helkkää	634
вя̄ййкшэ			вёййк	hämärä (s.)	730
нубпшэ	нубпай	itsepäinen	нубпь	itsepäisyys	909
поагшэ	поагкяй	itsepäinen	па̄гк	itsepäisyys	920
вагкшэ	вагкай	kauhistunut	вагк	kauhu	
лэ̄ншэ	лэ̄ннай	nuhainen	лэ̄нн	nuha	
мӯдтшэ	мӯдтай	pilaantunut	мӯдт	pilaantuneisuus	

* Esiintyy myös lähteessä KKLS

LIITE 3. VAIN YHDESTÄ KIELESTÄ TAVATUT 3-TAVUISET SENSIIVIJOHDOKSET

Eteläsaame: *råvnasjidh* (*rovnegs* ’outo’), ? *saajrasjidh* (? *saejrie* ’kipeä’)* (yht. 1–2)

Piitimensaame: *liejdahit* (*läjjdo* ’pitkäveteinen’) (yht. 1)

Luulajansaame: G *gárgahit* (*gárga* ’kitkerä’), *gávkahit* (*gávkas* ’pitkäveteinen’), *gånstahit* (*goansstá* ’konstikas’), G *gåntsahit* (*gåntsas* ’kummallinen’) (yht. 4)

Pohjoissaame: *árggášit* (*árgi* ’arka’), *árkkášit* (*árki* ’kurja’), *basášit* (*bassi* ’pyhä’), *boarkkašit* N (*boarka* ’äkkipikainen’), *buhtášit* (*buhtis* ’puhdas’), *buolašit* (*buolaš* ’pakkanen’), *buošášit* (*buošši* ’kiukkuinen’), *čalašit* N (*čalas* ’kova ja sileä’), *čavggašit* N (*čavgat* ’tiukka’), *čáhpašit* N (*čáhppat* ’musta’), S2 *čeahpášit* (*čeahppi* ’taitava’), *čieŋašit* (*čieŋal* ’syvä’), *čoaskkášit* (*čoaskkis* ’kylmä’), *čuovggášit* (*čuovgat* ’valoisa’), *dáiggášit* (*dáigi* ’taikina’), *dearvvašit* (*dearvvaš* ’terve’), *deašášit* (*deašši* ’heiveröinen’), *dilssášit* (*dilsi* ’ponneton’), *dimášit* (*dimis* ’pehmeä’), *divttášit* (*divttis* ’tiivis’), *doavkkášit* (*doavki* ’tyhmä’), *duolvvašit* (*duolvvas* ’likainen’), *erddošit* N (*erdui* ’ärtyisä’), *falášit* (*falli* ’nopea’), *fávrrošit* (*fávru* ’kaunis’), *funošit* N (*fuotni* ’huono’), S2 *galddašit* (*galda* ’pölkky’, *galdii* ’paksu(runkoinen)’), *gáiggašit* N (*gáiggas* ’typerä’), *gavžžašit* (*gavžžas* ’takakenoinen’), *gazášit* N (*gahci* ’saita’), *gámášit* N (*gámis* ’tumma’), *gáržžášit* (*gárži* ’ahdas’), *goalkkášit* (*goalki* ’tyyni’), *goasttášit* (*goasttis* ’ylivuotinen’), *goavášit* (*goavvi* ’ankara’), *guorbbašit* (*guorbbas* ’karvakulu’), *hávskkášit* (*hávski* ’hauska’), *hiđášit* N (*hiđis* ’hidas’), *hiljášit* (*hillji* ’rauhallinen’), *hivllášit* (*hivli* ’ohut’), *imašit* N (*imaš* ’ihmeellinen’), *jaŋašit* (*jaŋas* ’rutikuiva’), *jálddošit* (*jáldu* ’vilpoinen’), *jálošit* N (*jállu* ’rohkea’), *keamppašit* (*keampa* ’tyylikäs’), *lasášit* N (*lassi* ’lisä’), *lášššášit* N (*lášši* ’laiha’), *lávttašit* (*lávttas* ’kostea’), *livttášit* (*livttis* ’sileä’), *loanášit* (*loatni* ’veltto’), *loavddášit* (*loavdi* ’leppoisa’), *lojášit* (*lodji* ’kesy’), *maŋášit* (*maŋŋit* ’myöhäinen’), *njáđášit* (*njáđđi* ’matalalaitainen’; vrt. In *njiäđášiđ*), *njáiggošit* (*njáigu* ’vätys’), *njoazášit* (*njoahci* ’hidas’), *oanášit* (*oatni* ’lyhyt’) *ološit* (*ollu* ’paljo’), *rahpášit* (*rahpas*, *rahppái* ’ylhäältä väljä’), *ráhpášit* (*ráhpis* ’kivikkoinen’), *rievttašit* (*rievttęs* ’oikea’), *romášit* (*ropmi* ’ruma’), S2 *skavžžašit* (*skavžžas* ’kenottava, pysty’), *smáittášit* (*smáiti* ’heiveröinen’), *smirošit* (*smierru* ’hauras’), *snoakkášit* (*snoakkis* ’niukka’), *suivvašit* N (*suivat* ’pitkästyttävä’), *šiervvášit* N (*šiervi* ’hento’), *šluohkášit* (*šluohkis* ’katala’), *valjášit* (*vallji* ’runsaus’), *váttášit* (*váttis* ’vaikea’), *vávašit* (*vávvu* ’kumma’), *veahtašit* N (*veahtas* ’suolaton’), *vierášit* (*vieris* ’vieras’), *viiddášit* N (*viiddis* ’laaja’), *viissášit* (*viissis* ’viisas’), *vilddošit* (*vilddus* ’vuolas’), *vuorašit* (*vuoras* ’vanha’), *vuorjjášit* (*vuorji* ’harva’), S2 *vuovdášit* (*vuovdi* ’myyjä, innokas myymään’) (58 + 18 N + 4 S2 = yht. 80)

Inarinsaame: *komášiđ* (*komme* ’kumma’), *kumášiđ* (*kuṃe* ’kuuma’), *luátášiđ* (*luátis* ’väljä’), *miäđhášiđ* (*metki* ’hidasjalkainen’), *njiäđášiđ* (*njeeđi* ’matalalaitainen’; vrt. Po *njáđášit*), *piäivášiđ* (*peivi* ’päivä’), *ruákášiđ* (*ruok'kâd* ’rohkea’), *uinášiđ* (*uine* ’ujo’) (yht. 8)

Kildininsaame ks. Liite 2 (yht. 123)

* Eteläsaamen verbin *saajrasjidh* ’pitää liian hyvänä johonkin tarkoitukseen’ suhde oletettavaan kantasanaan (*saejrie* ’haava; kipeä’ tai *saejries* ’haavoittunut’) ei ole yksioikoinen. Lähtökohtana saattaisi olla ajatus, että jonkin hyvän haaskaaminen vähäarvoiseen tarkoitukseen tekee henkisesti kipeää; verrattakoon vastaavasti verbiin *sååjresjidh* ’kadehtia’ ja adjektiiviin *sååjrehke* ’kipeä, särkevä’ (< **sājru-*). Toisaalta verbi *saajrasjidh* voi olla kuitenkin pikemmin johdos vanhassa ruotsinsaamen kirjakielessä esiintyvästä sanasta *saire* ’saita, kitsas’ (LÖ:ssä myös asut *saires* ja *sairok*), joka lienee inarinsaamen adjektiivin *säiđi* ’id.’ tavoin lainaa suomen *saita*-sanan heikkoasteisista taivutusmuodoista (murt., vanh. *saiδan* ym.). Tällöin kyse ei siis olisi lainkaan sensiivijohdoksesta vaan johdos osoittaisi kantasanan tarkoitteena toimimista, ’kitsastelua’, mikä on *-š*-johtimen eräs toinen yleinen funktio.

Uumajansaame: *duvralussjat* (*duvras* ’kallis’)

Piitimensaame: *baskalussjat* (*basske* ’ahdas’), *binnalussjat* (*bin'ná* ’vähäinen’), *divralussjat* (*divras* ’kallis), *giedtsalussjat* (*gädtse* ’kapea’), *guhkalussjat* (*guhkke* ’pitkä’), *hejulussjat* (*hiedjo* ’köyhä’), *ienalussjat* (*iedna* ’paljon’), *ånålussjat* (*ådne* ’lyhyt’), *uhtsalussjat* (*uhttse* ’vähän’), *unnalussjat* (*un'ne* ’pieni’), *vánalussjat* (*vádne* ’niukka’), *vastalussjat* (*vasste* ’ruma’)

Luulajansaame: *agálussjat* (*ahket* ’ikävä’), *amálussjat* (*amás* ’vieras’), *áralussjat* (*árak* ’varhainen’), *binnalussjat* (*bin'ná* ’vähäinen’), *divralussjat* (*divras* ’kallis’), *ebdalussjat* (*iebdes* ’heiveröinen’), *enalussjat* (*enas* ’paljon’) , *gávkalussjat* (*gávkas* ’ikävä’), *guhkalussjat* (*guhkke* ’pitkä’), *gåntsalussjat* (*gåntsas* ’kummallinen’) , *imálussjat* (*imáj*/*imálasj* ’outo’), *lås(s)ålussjat* (*låssåt* ’raskas’), *nevralussjat* (*nievrre* ’huono’), AK *nuoralussjat* (*nuorra* ’nuori’), *sájgalussjat* (*sájgge* ’saita’), *skádalussjat* (*skádá* ’vahinko’), *stuoralussjat* (*stuorre* ’suuri’), *surgalussjat* (*surggat* ’surullinen’), *suttalussjat* (*suddo* ’synti’), *uhtsalussjat* (*uhttse* ’pieni’), *unnalussjat* (*un'ne* ’pieni’), *vájvalussjat* (*vájvve* ’vaivalloinen’), *vastalussjat* (*vasste* ’ruma’)

Pohjoissaame: *fuotninšakšit* (*fuotni* ’huono’), *garasšakšit* (*garas* ’kova’), *herskosšakšit* (*herskostallat* ’herkutella’, *herskui* ’herkutteleva’), *ruoinnasšakšit* (*ruoinnas* ’laiha’) || *aṣẹhuššat* (*aṣẹhaš* ’ohut’), *ahkidušša*t (*ahkit* ’ikävä, pitkäveteinen’)

Inarinsaame: *muččâdšukšâđ* (*muččâd* ’kaunis’), *sevŋâdšukšâđ* (*sevŋâd ~ siävŋád* ’pimeä’), *viššâlšukšâđ* (*viššâl* ’ahkera’), *arvâssukšâđ* (*aarvâs* ’antelias’), *puuriššukšâđ* (*puurič ~ puuriš* ’ahmatti’), *uánehšukšâđ* (*uánihâš* ’lyhyt’), *siälgáášukšâđ* (*siälgáá* ’liian loiva’), *irgáášukšâđ* (*irgáá* ’kosiskeleva’), *kimmáášukšâđ* (*kimmáá* ’kiimainen’), *sálttáášukšâđ* (*sálttáá* ’suolainen’), *skippiišukšâđ* (*skippii ~ skiivâs* ’laiha’), *tuolviišukšâđ* (*tuolvii* ’likainen’) || *omâšukšâđ ~ omâšuššâđ* (*oomâs* ’outo’), *poskâdšukšâđ ~ poskâšuššâđ* (*poskâd* ’ahdas’) || *njuoskâšuššâđ* (*njuoskâs* ’raaka’), *asašuššâđ* (*assaa* ’paksu’), *kumášuššâđ* (*kuṃe* ’kuuma’), *puolâšuššâđ* (*puolâš* ’pakkanen’)

Koltansaame: *â'ǩǩõõššâd* (*â'ǩǩed* ’ikävä’), *jâkstõõššâd* M (*jaakkâs*, **jiâkkâs* ’outo’), M *occnjõõššâd* (*occnjaž* ’vähäinen’), *õõmtõõššâd* (*õõmâs* ’kumma’)

Kildininsaame: *мыфтушшэ* (*тыввь* ’ihme’, KKLS 585), *ǩeäpsmõõššad* (*ǩie'ppes* ’kevyt’ KKLS 849), *puärsõõššad* (*puä´*res ’vanha’ KKLS 933), *jiâksõõššad* (*jaakkâs*, **jiâkkâs* ’outo’ KKLS 825), *veârsõõššad* (*vie'res* ’vieras’ KKLS 746, 961).

Vanhat sanakirjat: johdokset ja kantasanat esitetään lähteen mukaisessa asussa, merkitykset suomennettuina

Friis 1887: *akkedšokšat* (*akked* ’ikävä’), *amašavšet = amašet* (*amas* ’vieras’), *angeršavšat*, (*anger* ’innokas’), *apparšavšet* (*appar* ’iso’), *arkašavšet = arkašet* (*arkke* ’kurja’), *asašavšat = -šokšat = asašet* (*assai* ’paksu’), *bakašavšat = bakašet* (*bakas* ’kuuma’), *divrašavšet = divrašet* (*divres* ’kallis’), *dærvašavšet = dærvašet* (*dærvas* ’terve’), *fastašavšet = fastašet* (*faste* ’ruma’), *fuonošavšet = fuonošet* (*fuonos* ’huono’), *gălbmašokšat = gălmašet* (*gălmas* ’kylmä’), *gărašavšat = gărašet* (*găras* ’kova’), *gasašokšat = gasašet* (*găssag* ’paksu’), *gavkašokšat = gavkašet* (*gavkas* ’ikävystyttävä’), *gāvvelšavšat = -šokšat* (*gāvvel* ’viekas’), *giđašavšat = giđašet* (*giđđa* ’kevät’), *gukkašokšat = gukkašet* (*gukke* ’pitkä’), *guorossokšat = guorossšet* (*guoros* ’tyhjä’), *gæppašokšat = gæpašet* (*gæppad* ’kevyt’), *lojašokšat = lojašet* (*logje* ’säyseä’), *nævrašokšat = nævrašet* (*nævrre* ’paha, kurja’), *oanašokšat = oanašet* (*oadna* ’lyhyt’), *ovdušokšat = ovdušet*, (*oavddo* ’ihme’), *unnašokšat = unnašet* (*unne* ’pieni’), *ænašokšat = ænašet* (*ænas* ’enin’)

Leem 1768: *amashjashjam* (*amas* ’vieras’), *buorashjashjam* (*buorre* ’hyvä’), *fastashjashjam* (*fasste* ’ruma’), *garashjashjam* (*garas* ’kova’), *loijashjashjam* (*loigje* ’säyseä’), *mokkashjashjam* (*mokke* ’mutka, petkutus’)

Lindahl & Öhrling 1780: *wuorahaksjet* (*wuoras* ’vanha’), ? *tåbdahaksjet* (*tåbdet* ’tuntea’, *tåbdos* ’tunnettu’)

Plantinus: *vastaiaxet* (vrt. LÖ *waste* ’ruma’)

Creating a corpus for Kven, a minority language in Norway

Pia Lane, Kristin Hagen, Anders Nøklestad and Joel Priestley
University of Oslo

Abstract

Language documentation, including the development and use of corpora, is frequently linked to revitalisation. This is also the case for the Kven language, a Finnic minoritised language, traditionally spoken in the two northernmost counties of Norway. Kven is a recognised minority language in Norway, protected by the European Charter for Regional or Minority Languages. This status led to increased efforts to document Kven, including the development of the Ruija Corpus, consisting of recordings of interviews in Kven. The corpus was an important tool for the standardisation of Kven. In this article we describe how the corpus was developed and account for search functions, including a discussion of the limitations of the corpus. We also discuss the role of corpora and other online tools for language revitalisation, with a particular focus on the standardisation of Kven and conclude by reflecting on how expertise also resides with the speakers of an endangered language and that they have a right to be involved in efforts of language documentation and revitalisation.

Keywords: Corpus linguistics, revitalisation, minority language, Kven

1. Introduction

Across the world, a large number of languages are in the process of being standardised, following a longstanding tradition within anthropological linguistics, linguistic typology and language documentation. Scholars describe and document languages, and we have sophisticated means for data analysis, for developing grammars and dictionaries, and recently also for building large electronic corpora. One such corpus is the Ruija Corpus, a speech corpus from Kven and Finnish-speaking areas in Northern Norway. In this article, we first provide some information on the Kven people and their language and outline the development of the Ruija Corpus and describe search functions, including a discussion of the limitations of the corpus, due to the lack of grammatical annotation. We then discuss the role of corpora and other online tools for language revitalisation, with a particular focus on the standardisation of Kven and conclude by reflecting on how expertise also resides with the speakers of minoritised language and that they have a right to be involved in efforts of language documentation and revitalisation.

2. Background – Kven language

The Kven language is a Finnic minoritised language, traditionally spoken in Troms and Finnmark, the northernmost counties of Norway, see the map in figure 1. The Arctic region has been multilingual for centuries, and from the beginning of the 18th century, people from Finnish-speaking areas, in what today are the northern parts of Sweden and Finland, settled along the coast of Northern Norway; some of them settled before the current national frontiers were drawn (Sundelin 1998). This group of people and their descendants are called Kven or Norwegian Finns, and particularly in the coastal areas there is a long tradition of trade and intermarriage with the Sámi population. When the idea of Norway as a nation state got foothold in the 19th century, a monolingual and homogenous nation came to be one of the cornerstones of the idea of the Norwegian nation state, and assimilatory and even oppressive policies were directed towards the Kven and Sámi populations. These policies are referred to as the Norwegianisation process, and many of these were directed towards language, and regulations limited the use of Kven and Sámi in schools and sale of land in the northern areas to people with knowledge of Norwegian (Pietikäinen et al. 2010). The Kven were seen as a 'national problem' because of their position as a border minority, "Russia's

 Nordlyd 46.1: 159–170, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6345

foothold in Western Europe" (Niemi 1995: 196); hence, the assimilation policies were particularly prominent in the exposed border areas (Niemi 2003). In tandem with general processes of modernisation, these policies contributed to an extensive language shift from Kven to Norwegian, and by the 1960s most Kven children spoke Norwegian only. Norway's ratification of the European Charter for Regional or Minority Languages (under the auspices of the Council of Europe) in 1993 subsequently led to the recognition of Kven as a language in 2005, and not primarily a dialect of Finnish (Lane 2011).

In Finland, Finnish was given official status in 1863 and was developed into a language of education and administration. Vocabulary and grammatical structures from the Eastern Finnish dialects were included, and neologisms were created in order to replace Swedish loanwords (Latomaa and Nuolijärvi 2005), processes which influenced the Finnish dialects in Finland, whereas the Kven had left before this standardisation took place. Because the Kven were not in Finland during the standardisation process, their language developed differently, particularly in terms of vocabulary. Kven and Finnish are agglutinating languages with rich inflectional morphology, but the largest difference between Kven and standard Finnish is lexical due to old Swedish borrowings that have been retained in Kven and newer borrowings from Norwegian. Kven and Finnish are mutually intelligible, though Finnish speakers understand Kven better than Kven speakers understand Finnish because Finns learn Swedish in school and therefore understand older Swedish and Norwegian borrowings in Kven, whereas Kven speakers in most cases do not understand the Finnish equivalents of these borrowings (Lane 2016). Self-identification is also one of the criteria for the language – dialect distinction: the Kven speakers identify themselves as Norwegian and not Finnish (Hyltenstam and Milani 2003).

Figure 1. Map Troms and Finnmark County. Source: Kartverket – Norwegian Mapping Authority (Attribution 4.0 International (CC BY 4.0) © Kartverket https://www.kartverket.no/en).

3. How the corpus came about

The fourth International Polar Year (IPY 2007–2008) was the largest global research initiative to be carried out for 50 years, and approximately 50 000 researchers and language technologists from 60 countries participated. The Research Council of Norway issued a Call for Proposals, which resulted in 29 funded projects. One of these was the Linguistic and Cultural Heritage Electronic Network (LICHEN) at the Department of Linguistics and Scandinavian studies (University of Oslo), directed by Pia Lane. The project was a part of a larger international cluster, whose aim was to create an electronic framework for the collection, management, online display, and exploitation of existing corpora of the languages of the northern circumpolar region. The Norwegian part of the LICHEN project involved carrying out a pilot project on the Kven language. The aim of the Norwegian LICHEN project was to digitise and transcribe old recordings of Kven, collect new data in order to test the tools being developed by the University of Oulu, and to make the recordings and transcriptions available for researchers. The main electronic system was not completed by the time the Norwegian project was concluded, so a separate speech corpus, the Ruija Corpus, was developed by Pia Lane in cooperation with the Text Laboratory at the Department of Linguistics and Scandinavian Studies, University of Oslo. Ruija (both in Kven and Finnish) refers to

Northern Norway, and the name was chosen for the corpus not to alienate speakers who prefer to refer to their language as Finnish (often modified as 'our Finnish' to distinguish it from standard Finnish).

As is often the case for minority language projects, there was limited funding and thus lack of resources to make an annotated corpus with linguistic information (such as lemmas and morphosyntactic tags). The PI of the Norwegian LICHEN project, Pia Lane, did not have a permanent position, which further limited the time available for corpus development. The corpus consists of recordings and transcriptions. When the corpus was initiated, there was no written standard for Kven, so there was no standard that could be used as a basis for the transcriptions. The recordings were transcribed by students at the University of Oulu and Mikael Voronov at the Kven Institute. The transcribers followed the same guidelines, but they were based at different institutions and worked with different supervisors who implemented the guidelines slightly differently. Consequently, there is some variation in the transcription conventions.

The corpus was launched in April 2010 with 76 hours of recordings, and data from two other projects added. These projects are *Identities in transition – a longitudinal study of language shift* and *Standardising minority languages – STANDARDS*, both financed by the Research Council of Norway and directed by Pia Lane. *Identities in transition,* a study of language shift in Bugøynes-Pykejä (a Kven community in Northern Norway) compared and contrasted interview data from the same individuals from 1975 and 2008, supplemented by interviews in Norwegian with younger speakers from the generation who had shifted to Norwegian. The STANDARDS project explores how intended users of a written standard of Kven relate to the standard and consists of interviews and video recordings of Kven speakers reading a text in Kven for the first time. The project source codes in the corpus reflect these projects: LICHEN = LI and KI, Identities in transition = id and sks,[1] SMS = STANDARDS. From 2023 the corpus will also contain interviews from the project *Voices of revitalisation,* which focuses on how new speakers experience the process of starting to speak Kven or Sámi. New speakers are individuals who have learned an indigenous or minoritised language in an educational setting, often as a part of revitalisation efforts, and who reclaim (start speaking) the language later, often at important transitional moments in life, such as when needing the language for work purposes or becoming a parent (O'Rourke and Pujolar 2015). The corpus therefore has data suited for analyses of grammar or phonology, including change over time as the main bulk of the corpus spans a period of more than 30 years. This timespan also allows for sociolinguistic studies exploring changes in language attitudes and identity construction over time.

4. Developing the corpus

The first version of the Ruija Corpus was presented in an older version of the search and postprocessing tool Glossa, developed at the Text Laboratory (Johannessen et al. 2008). The main idea behind Glossa was – and still is – to give researchers a user-friendly tool where they can concentrate on their research and do not need to learn advanced query languages or attend courses to use the tool.

This first version of Ruija contained nearly 430 000 tokens, 379 000 Kven and 50 300 Norwegian ones. There were 85 speakers in the corpus from 12 places. Some of the speakers were interviewed twice (in 1975 and 2009, as a part of the project *Identities in transition*).

The present version of the Ruija Corpus is converted to a new and even more user-friendly version of Glossa (Nøklestad et al. 2017, Søfteland et al. 2020).[2] The corpus is enriched with more speech data and in 2022 it contains almost 522 000 tokens from 12 places, with 109 speakers in total.

The main search page is very simple as figure 2 shows. You can search for one or more words in the Google-like search box in the middle of the page. Metadata categories are located on the left-hand side. The results are given as concordances. Figure 3 shows a search for "kveeni"'. 'Kveeni' is a Kven word that refers to both the Kven people and their language and may be used both as a noun and as an adjective. Above the metadata menu you can see how many speakers and words are included in the chosen selection. A click on "Show speakers" gives you a list of all speakers in the selection.

[1] sks refers to Suomalaisen Kirjallisuuden Seura (The Finnish Litterature Society) which provided the recordings from 1975.

[2] Glossa uses the IMS Open Corpus Workbench system for text search and MySQL as the metadata database. The server code is written in Clojure, a modern dialect of Lisp that runs on the Java Virtual Machine.

To the left of each search result there are two or three icons: one for playing the search result in a media player (figure 4) and one for showing the search result as a waveform (figure 5). Some recordings also have a button for video presentation. Clicking on the speaker's identifier shows all the metadata available for the speaker.

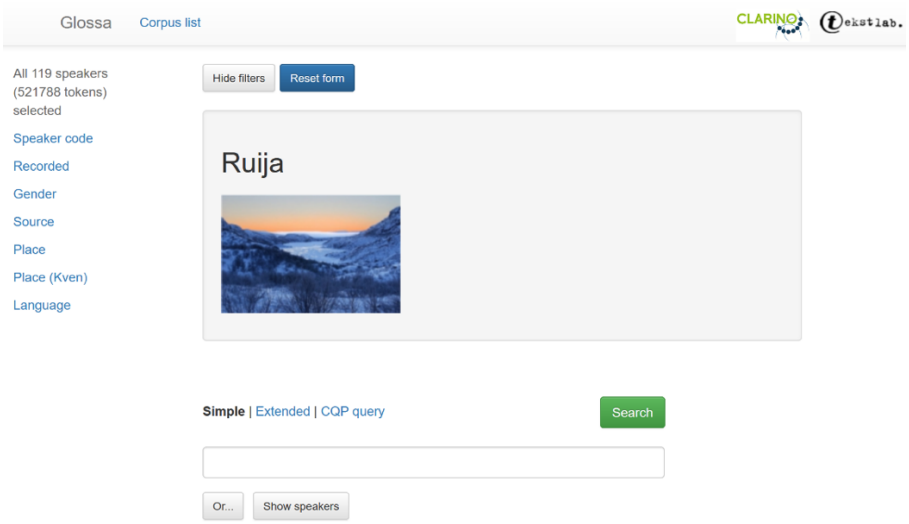


Figure 2. The main search page of the Ruija Corpus.

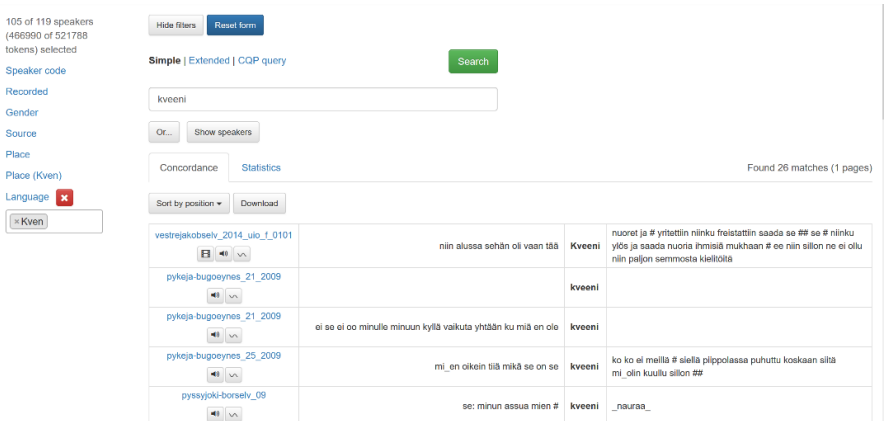


Figure 3. Example of simple search result: kveeni – 'Kven'

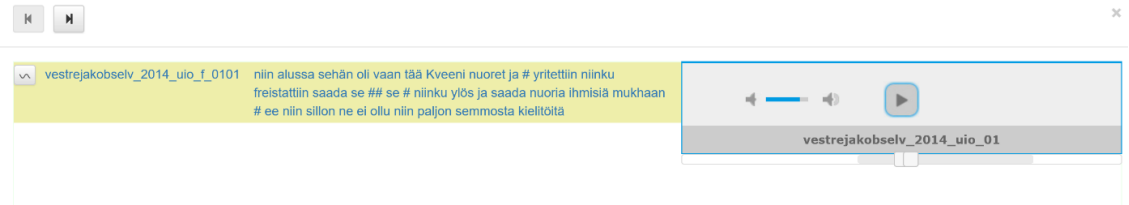


Figure 4. The search result in a media player. If you move the squares under the box left and/or right you get more context.

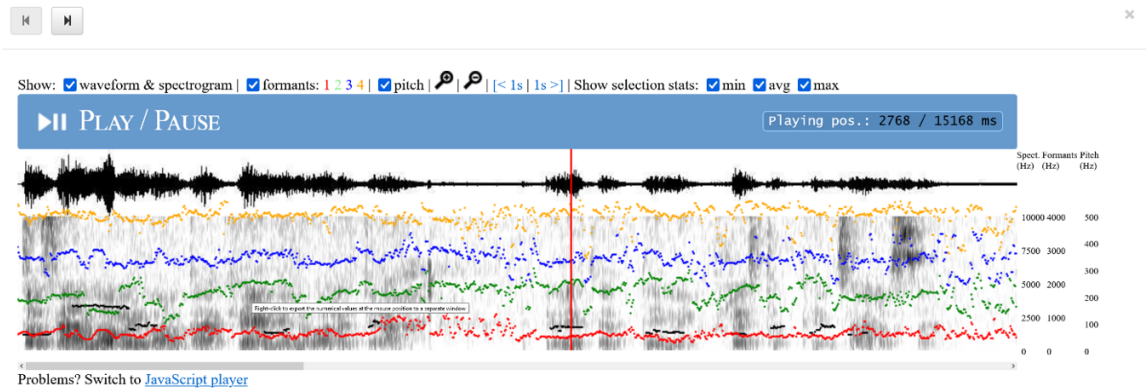


Figure 5. Example of search result as waveform.

Above the search box at the main search page there are two more search options: extended search and CQP search. If you know the query language of the Corpus Query Processor (CQP) search engine you can make your own advanced searches in the CQP box. If you want advanced searches by using menus and text fields, choose the *Extended* option, see figure 6. Here you can choose to search for the start of a word, the middle or the end. You can also search for the first or final word in a speech segment. Since the corpus is not lemmatised or part-of-speech tagged yet, the menu symbol on the left can only offer a box for excluding word forms.

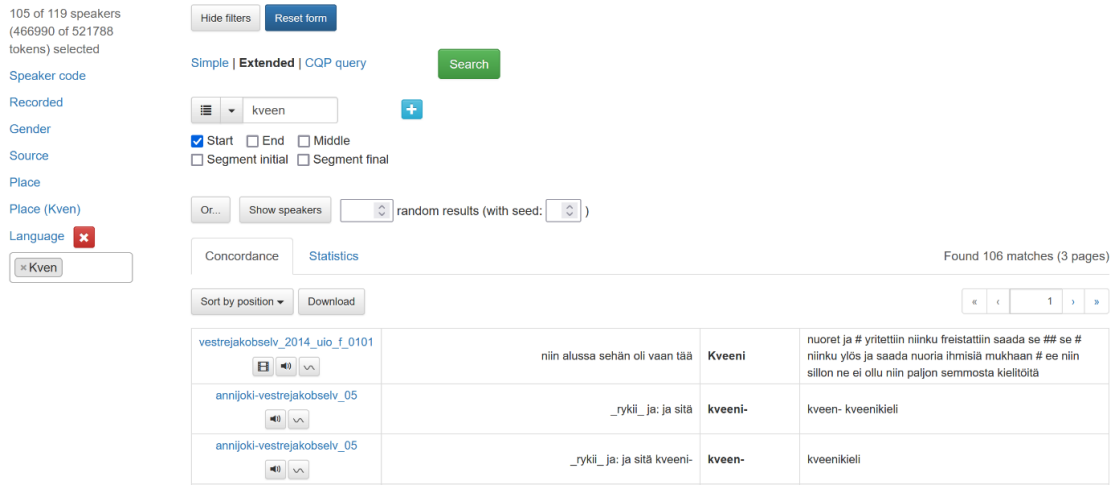


Figure 6. Example of extended search.

Figure 6 shows how to search for different word forms or inflected forms even if the corpus is not part-of-speech tagged. A search for "kveen" as the start of the word will show all words starting with "kveen" as figure 7 shows[3]. Here we have chosen to show the search result as a word list by clicking on "Statistics" below the search result. The search results are downloadable in various formats.

[3] The transcribers followed orthographic conventions, and the stem vowel of Kven was pronouced both as e or æ, transcribed as kveeni or kvääni, respectively. A search for word initial kv- would have yielded both forms.

Update stats Download: Excel Tab-separated Comma-separated

Count	Word form
27	kveenin
26	kveeni
7	kveenit
5	kveenien
4	kveeneiksi
3	kveeninkieli
3	kveenejä
3	kveen-
2	kveeniä
2	kveeniliitto
2	kveeni-
2	kveeni'asia
1	kveensk
1	kveenlan
1	kveenistä
1	kveenir_rahatki
1	kveenipuvun

Figure 7. The search result as a word list.

The Ruija Corpus is available for research, but due to the personal nature of the content in the corpus, the license is a restrictive, academic one. You have to send an application to the owner of the corpus, and after you have been granted access, you can login with Feide (the centralized identity management solution for the educational sector of Norway) or CLARIN (Common Language Resources and Technology Infrastructure).

5. Standardisation of Kven

As a consequence of the recognition of Kven, a standardisation process was initiated. The Kven Language Body, consisting of the Kven Language Council (advisory function) and the Kven Language Board (executive function) developed principles for the standardisation of Kven. They recommended that the standard should be close to Meänkieli (a closely related Finnic language in Sweden), that preference be given to forms common in several Kven dialects, that one should not aim to make the standard as removed from Finnish as possible, and that the standard should be based on standard Finnish orthography. There are some differences in grammatical structures between the Western and Eastern Kven varieties, as the settlement in the Eastern area occurred later and these areas in general had somewhat more contact with Finland (Lane 2017; Keränen 2018). An example is the interdental fricative /ð/, a phoneme that has been retained by some Kven speakers in Børselv-Pyssyjoki and is used by the writers from this village. For some it is a strong identity marker, and the Language Body therefore decided that in language regions where there is a need for additional letters to the Finnish orthography, such as *š* (alternatively sh) and *đ* / *ð*, these may be used.[4] In the Kven grammar the letter <đ> is used consistently throughout the grammar to represent /ð/ even though apart from a few Kven speakers in the Western, /ð/ has not been retained in Kven. For those who use /ð/, there is alternation between /t/ and /ð/, whereas the majority of the Kven dialects have alternation between /t/ and /Ø/,[5] see Lane (2016) for an analysis.

A grammar of Kven was published in Kven in 2014 and translated to Norwegian in 2017 (Söderholm 2014, 2017). The grammar was approved by the Kven Language Board, the decision-making body for the school norm of Kven. According to Evans and Dench (2006), the aim of a descriptive grammar is to capture and codify the essential structural features of a language, ideally collected as a part of a programme on language documentation, often based on a natural speech data, but sometimes supplemented by speaker acceptability judgements (p. 3). Language documentation has a broader aim than providing a grammatical

[4] In Finland, *š* is used when writing some foreign names.
[5] Ø is used to denote that there is an alternation between a consonant and zero (such as *pöytä – pöyän* 'table-NOMINATIVE - table-ACCUSATIVE).

description of a language; ideally such efforts should also document how the languages is *used*; thus, there is also an emphasis on cultural and social aspects of language (Austin 2020).

The Ruija Corpus was one of the sources for the writing of the grammar book (Söderholm 2014, 2017) in addition to other recordings and transcriptions of Kven dialect and written sources. Literature in Kven is still limited, so the author of the grammar drew primarily on novels published by Alf Nilsen-Børskog, from Børselv-Pyssyjoki in the Western dialect area, therefore the grammar is written in the Western variety of Kven while providing information about the Eastern varieties (Lane 2017). The Ruija Corpus was used both by the Kven Language Council and later by Söderholm as a base for the grammatical description of Kven as it contained recordings and transcriptions of interviews from all the core Kven areas, this allowing the council and Söderholm (who also was a member of the Kven Language Council) to map the key dialectal differences (see also Östman 2000 for a discussion on research ethics in minoritised and Indigenous contexts). Dealing with variation was one of the main challenges the Kven Language Body faced, because standardisation always entails reducing and abstracting away from diversity, as pointed out already by Milroy and Milroy (1999). Their aim was to develop a standard that could be used as a basis for developing teaching materials for learners of Kven in the educational system and for producing texts for people who spoke Kven. When considering the needs of the learners, there was a concern that a standard with too many options could impact learning negatively, whereas there also was a need to allow for enough variation for speakers of Kven to recognise 'their language' and identify with the standard. The standardisation of minority languages is an ambivalent process because it requires selecting particular forms over others — they generate and legitimise high varieties in minority languages as well as the structures to sustain their diffusion, potentially establishing linguistic standards that the language speakers themselves experience that they cannot meet (Costa, De Korne and Lane 2017). Consequently, minority language speakers are potentially faced with a double stigma (Gal, 2006; Lane 2011): their language falls short when measured against the official national language and in terms of meeting the standardised version of the minority language. The written standard may therefore be perceived by social actors as lacking both the authority and invisibility of a national language and the authenticity and legitimacy of the minority language (Woolard 2008; Lane 2015).

Currently, the most widely used version of the standard is the form closest to the one used by Söderholm (2014), though those who write Kven adapt their texts to the local context, a task for which the Ruija Corpus provides useful information. Even though the corpus is not lemmatised or part-of-speech tagged, searches for high-frequent words may still yield a substantial amount of information. One example is variation in the realisation of infinitive forms. Some Western Kven dialects have retained the proto-Finnic word final /t/ (alternation with or without /t/) whereas this is not the case for the Eastern dialect. In the absence of grammatical information, an option is to search for word forms. As almost all those who were interviewed were asked about languages they use, searches for *puhua* 'speak' would yield a lot of examples. As there is a considerable amount of phonological variation and morphophonological alternation, the best suited option is to search for the beginning of the word and then search for infinitive forms manually. Figure 8 is an example of the infinitive form of *puhua* 'speak' with word-final /t/: se oli semmonen häpy et ei pitänyt puhhut kvääanin kieltä # - it was such a shame that one should not speak Kven language # mm'[6]

Figure 8. Infinitive with word-final /t/.

The following figure shows the infinitive form of *puhua* without word-final /t/: mutta sehän oli on # eri aika nyt ## mutta sillon ei saanu puhua # suomea 'but then of course was # a different time now ## but then one wasn't allowed to speak # Finnish'.

[6] # = pause

Figure 9. Infinitive without word-final /t/.

This is of course a search method which is somewhat cumbersome and time consuming and also requires that the user has a fairly good knowledge of Kven dialects and grammar or at least a basic understanding of Finnish grammar, but this still allows us to get a basic overview of grammatical variation in Kven dialects. For researchers who wish to conduct sociolinguistic or historical studies, there are two main options for searches: one may either search for key words such as language, Kven, Finnish, Norwegian, school, war etc. in the actual corpus or search the full transcripts of each interview (there is a link to transcriptions on the Ruija Corpus main page). The latter would also allow for discourse analysis or narrative analysis.

6. Role of corpora in the revitalisation of minoritised languages

Language revitalisation may be seen as a part of Language Policy and Planning (LPP), a discipline which initially developed as a part of sociolinguistics and language-in-society studies and emerged as a field of study in the 1960s (Kaplan et al. 2000). Initially, the main efforts to plan the use and status of languages focused on national languages and nation building (Wright 2004), and language was seen as a static and delimited entity, an object which could be captured and codified. The overarching term in this period was Language Planning, with a focus on linguistic aspects and the use and status of language (Kloss 1967, Haugen 1972).[7] The structuralist period after WWII laid the foundations of what was to characterise LPP until the critical turn in the social sciences and humanities in the 1970s which brought a stronger focus on context and language use, even questioning core concepts such as language and native speaker (Lane 2015). *Language revitalisation* is commonly understood as community and individual efforts to maintain an indigenous or minoritised language or 'giving new life and vigour to a language that has been decreasing in use (or has ceased to be used altogether)' (Hinton, Huss and Roche 2018: xxi). For such efforts language documentation is important, which for Kven, the Ruija Corpus is a key part. Trond Trosterud has been a key contributor to the development of other online resources (available to the general public), such as an online dictionary (Trosterud 2019) and a morphological analyser for Kven developed by Giellatekno (Trosterud et al. 2017), the Research group for Sámi language technology, at UiT The Arctic University of Norway. The dictionary is written in the Børselv-Pyssyjoki dialect, but the analyser contains morphophonological information from different dialects. Thus, the dictionary recognises words from different dialects, such that a search for *lukea* 'read', common in the Eastern varieties, yields a Norwegian translation, but provides the Western infinitive form with a word final <t> *lukkeet*:

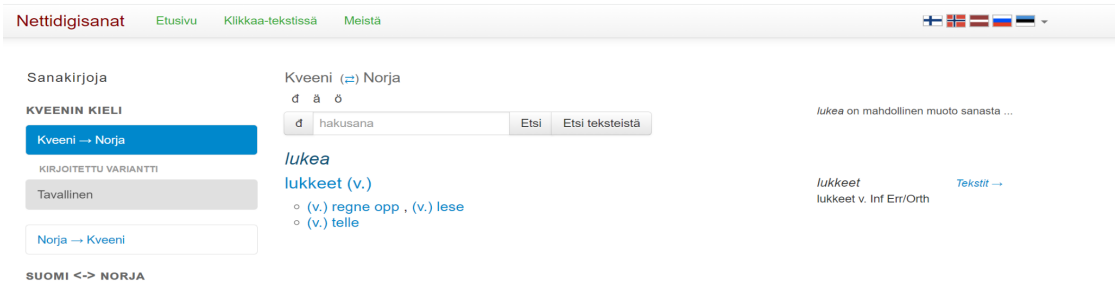


Figure 10. Example of entry from the Kven online dictionary: lukea – 'to read'.

[7] The term 'Language Planning' frequently is attributed to Haugen, but he mentions that Weinreich used the term Language Planning as a title for a seminar in 1959 (Haugen 1972: 209).

Ideally, such an online dictionary should also provide forms common in other Kven dialects, but as often is the case when documenting and standardising minoritised languages, human and monetary resources are limited; therefore, tools are built step-by-step and based on available resources. All forms of language documentation and standardisation, including making corpora, dictionaries and grammars, are carried out with a user in mind, though actors are aware of this to varying degrees, and decisions to include, and thereby exclude, some grammatical forms are not a purely linguistically based choice. The choices made by researchers and language planners may constrain future actions of intended users as these might not recognise the way they speak or lack resources or knowledge to use the developed tools. More importantly, corpora, dictionaries and grammars also prepare efforts to revitalise a minority language as such tools provide the basis for educational materials, literature and more visibility for minority languages in public space. This has indeed been the case for Kven as textbooks and a grammar for the educational system have been developed, an annual New Year speech in Kven is aired by the national Norwegian Broadcasting Corporation, street signs in Kven are introduced in the northern part of Norway giving visibility to traditional Kven place names, and the Norwegian nation state has got a Kven name (in addition to Norwegian and Sámi), namely *Norja*. There is an urgent need for more arenas for the use of Kven, both within and outside the educational system. Hinton (2018: 460) reminds us that:

> The biggest hurdle for both native speakers and language learners is to actually start using the language on a daily basis. For endangered languages, this is a major challenge. Just as elders in a community that has undergone language shift cease to use the language they grew up with because most of the community doesn't know it, so do second-language learners find themselves without interlocutors.

In spite of limited resources, dedicated researchers and language technologists have managed to develop a range of resources that have been used and will continue to be used in the process of revitalising Kven. What now remains is to continue developing the corpus, not only by including new material but also by developing the Ruija Corpus into a morphologically tagged corpus.

7. Conclusion: future prospects

Language documentation and revitalisation used to rely extensively on the role of academic experts, but there has been a shift in the field recognising that expertise also resides with the speakers of an endangered language and that they have a right to be involved in and shape these processes (Hill 2002) as a part of "a larger effort by a community to claim its right to speak a language and to set associated goals in response to community needs and perspectives" (Leonard 2017: 19). This was also the case for the documentation of Kven as the Kven Language Council had Kven members and all the members of Language Board, which held the executive function, were Kven speakers. When documenting minoritised languages, linguists work in tandem with members of local communities and participate in community efforts to sustain languages. Such participation in turn influences the balance of power and opens up space for new types of knowledge, as outlined by Eira (2007), see also Lane and Makihara (2017) for a discussion.

When the Ruija Corpus was created, such concerns were less prominent in our research field, and consequently, the corpus is only available to researchers and those who are interviewed are anonymised. This is a twofold challenge: the corpus is not available to speakers and learners of Kven, and the expertise and knowledge of language and culture by those interviewed is not acknowledged because according to ethical regulations when the creating the corpus began in 2007, their anonymity had to be ensured. While the mind-set of the academic community has undergone a profound change striving to recognise the knowledge, voices, perspectives and expertise of the speakers and communities we are working with, there are still many hurdles to overcome. A future endeavour for the Ruija Corpus will be to find ways of making at least parts of the corpus available also for users outside the traditional academic community, thus recognising that knowledge and ownership to data do not reside within academia only.

References

Austin, Peter. 2020. Language documentation and revitalisation. In *Revitalizing Endangered Languages: A Practical Guide*, edited by Justyna Olko and Julia Sallabank, pp. 199–219. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108641142.014.

Costa, James, Haley De Korne and Pia Lane. 2017. Standardising minority languages: Reinventing peripheral languages in the 21st century. In *Standardizing Minority Languages: Competing Ideologies of Authority and Authenticity in the Global Periphery*, edited by Pia Lane, James Costa and Haley De Korne, pp.1–23. Routledge, New York.

Eira, Christine. 2007. Addressing the ground of language endangerment. In *Working Together for Endangered Languages: Research Challenges and Social Impacts – Proceedings of Foundation for Endangered Languages Conference XI Kuala Lumpur October 26–28* 2007, edited by Maya K. David, Nicholas Ostlerand Ceasar Dealwis, pp. 82–90. Foundation for Endangered Languages.

Evans, Nicholas and Alan Dench. 2006. Introduction: Catching language. In *Catching Language: The Standing Challenge of Grammar Writing*, edited by Felix Ameka, Alan Dench and Nicholas Evans, pp. 1–39. Mouton de Gruyter, Berlin, New York.

Gal, Susan. 2006. Contradictions of standard language in Europe: Implications for the study of publics and practices. *Social Anthropology*: 14(2), 163–181. https://doi.org/10.1111/j.1469-8676.2006.tb00032.x.

Glossa: https://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/glossa/index.html

Haugen, Einar. 1972. *The ecology of language: Essays by Einar Haugen. Selected and introduced by Anwar S. Dil*. Stanford University Press, Stanford.

Hill, Jane. 2002. "Expert rhetorics" in advocacy for endangered languages: who is listening, and what do they hear? *Journal of Linguistic Anthro*pology 12(2): 119–133. https://doi.org/10.1525/jlin.2002.12.2.119

Hinton, Leanne. 2018. Approaches to and strategies for language revitalization. In *The Oxford Handbook of Endangered Languages*, edited by Kenneth Rehg and Lyle Campbell, pp. 443–465. Oxford University Press, New York. https://doi.org/10.1093/oxfordhb/9780190610029.013.22.

Hinton, Leanne, Leena Huss and Gerald Roche. 2018. Language revitalization as a growing field of study and practice. In *The Routledge Handbook of Language Revitalization,* edited by Leanne Hinton, Leena Huss and Gerald Roche, pp. xxi-xxx. Routledge. New York.

Hyltenstam, Kenneth and Tommaso M. Milani. 2003. *Kvenskans Status: Rapport for Kommunal- og regionaldepartementet og Kultur- og Kirkedepartementet i Norge*. Oslo.

IMS Open Corpus Workbench: http://cwb.sourceforge.net.

Johannessen, Janne Bondi; Nygaard, Lars; Priestley, Joel; Nøklestad, Anders. 2008. Glossa: a multilingual, multimodal, configurable user interface. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Daniel Tapias, pp. 617–621. European Language Resources Association (ELRA), Paris.

Kaplan, Robert, Richard Baldauf, Anthony Liddicoat, Pauline Bryant, Marie-Thérèse Barbaux and Martin Pütz. 2000. Current issues in language planning, *Current Issues in Language Planning*, 1(1): 1–10. https://doi.org/10.1080/14664200008668003.

Keränen, Mari. 2018. Language maintenance through corpus planning – the case of Kven. *Acta Borealia*, 35(2): 176–191. https://doi.org/10.1080/08003831.2018.1536187.

Kloss, Heinz 1967Bilingualism and nationalism. *Journal of Social Issues*, 23(2), 39–47. https://doi.org/10.1111/j.1540-4560.1967.tb00574.x

Kven online dictionary, Nettidigisanat https://sanat.oahpa.no/

Latomaa, Sirkku and Pirkko Nuolijärvi. 2005. The language situation in Finland. In *Language Planning and Policy in Europe, Vol. 1. Hungary, Finland and Sweden*, edited by Robert B. Kaplan and Richard B. Baldauf, pp. 125–232. Multilingual Matters, Clevedon.

Lane, Pia. 2011. The birth of the Kven language in Norway: Emancipation through state recognition. *International Journal of the Sociology of Language* 209: 7–74. https://doi.org/10.1515/ijsl.2011.021.
Lane Pia. 2016. Standardising Kven: Participation and the role of users. *Sociolinguistica 30*: 105–124. https://doi.org/10.1515/soci-2016-0007.
Lane, Pia. 2015. Minority language standardisation and the role of users. *Language Policy* 14, 263–283. https://doi.org/10.1007/s10993-014-9342-y
Lane, Pia. 2017. Language standardisation as frozen mediated actions – the materiality of language standardization. In *Standardizing Minority Languages: Competing Ideologies of Authority and Authenticity in the Global Periphery*, edited by Pia Lane**,** James Costa and Haley De Korne, pp. 101–117. Routledge, New York. https://doi.org/10.4324/9781315647722.
Lane, Pia and Miki Makihara. 2017. Indigenous peoples and their languages. In *The Oxford Handbook of Language and Society*, edited by Ofelia García, Nelson Flores and Massimilano Spotti*,* pp. 299–230. Oxford University Press, New York. https://doi.org/10.1093/oxfordhb/9780190212896.013.7.
Leonard, Wesley. 2017. Producing language reclamation by decolonising 'language'. *Language Documentation and Description* 14: 15–36. http://www.elpublishing.org/PID/150.
Milroy, James and Leslie Milroy. 1999. *Authority in Language: Investigating Standard English*. Routledge, London.
Niemi, Einar. 1995. The Finns in northern Scandinavia and minority policy. In *Ethnicity and Nation Building in the Nordic World*, edited by Sven Tägil*,* pp. 145–178. Hurst and co, London.
Niemi, Einar. 2003. Regimeskifte, innvandrere og fremmede. In *Norsk innvandringshistorie. I nasjonalstatens tid 1814–1940,* edited by Einar Niemi, Jan Eivind Myhre and Knut Kjeldstadli, pp. 11–47. Pax forlag, Valdres.
Nøklestad, Anders, Kristin Hagen, Janne Bondi Johannessen, Michal Kosek and Joel Priestley. 2017. A modernised version of the Glossa corpus search system. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa),* edited by Jörg Tiedemann and Nina Tahmasebi, pp. 251–254. Association for Computational Linguistics, Gothenburg.
O'Rourke, Bernadette and Joan Pujolar. 2015. New speakers of minority languages: the challenging opportunity – Foreword. *International Journal of the Sociology of Language* 231: 1–20. https://doi.org/10.1515/ijsl-2014-0029.
Pietikäinen, Sari, Leena Huss, Sirkka Laihiala-Kankainen, Ulla Aikio-Puoskari and Pia Lane. 2010. Regulating multilingualism. *Acta Borealia*: 27(1): 1–23. https://doi.org/10.1080/08003831.2010.486923
Ruija Corpus: https://tekstlab.uio.no/glossa2/ruija
Sundelin, Egil. 1998. Kvenene – en nasjonal minoritet i Nord-Troms og Finnmark? In *Kvenenes historie og kultur*, edited by Helge Guttormsen, pp. 35–48. Nord-Troms historielag, Skjervøy.
Söderholm, Eira. 2014. *Kainun Kielen Grammatikki*. Suomalaisen Kirjallisuuden Seura, Helsinki.
Söderholm, Eira. 2017. *Kvensk Grammatikk*. Cappelen Damm Akademisk, Oslo.
Søfteland, Åshild, Anders Nøklestad, Joel Priestley and Kristin Hagen. 2020. Glossa som forskningsverktøy. Hva folk søker etter og hva resultatene brukes til. *Oslo Studies in Language*: 11(2): 449–464. https://doi.org/10.5617/osla.8512.
Trosterud, Sindre Reino, Trond Trosterud, Anna-Kaisa Räisänen, Leena Niiranen, Mervi Haavisto and Kaisa Maliniemi. 2017. A morphological analyser for Kven. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pp. 76–88, St. Petersburg, Russia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-0608.
Trosterud, Trond. 2019. Kva bruker vi minoritetsspråksordbøker til? Ein studie av brukarloggane for tolv tospråklege ordbøker. *LexicoNordica* 26: 177–198.
Woolard, Katheryn. 2008. Language and identity choice in Catalonia: The interplay of contrasting ideologies of linguistic authority. In *Lengua, Nación e Identidad: La Regulación del Plurilingüismo en España y América Latina*, edited by Kirsten Siiselbeck, Ulrike Mühlschlegel, and Peter Masson, pp. 303–323. Vervuert, Frankfurt am Main /Iberoamericana, Madrid.

Wright, Sue. 2004. *Language Policy and Language Planning: From Nationalism to Globalisation*, Palgrave Macmillan, Basingstoke.

Östman, Jan-Ola. 2000. Ethics and appropriation – with special reference to Hwalbáy. In *Issues of Minority Peoples*, edited by Frances Karttunen and Jan-Ola Östman, pp. 37–60. Department of General Linguistics, University of Helsinki.

Anarâškielâ postpositioi *pelni* já *piälán* čäällim sierâ já oohtân tievâdâsâinis SIKOR-tekstâčuágálduvâst

Petter Morottaja
Oulu ollâopâttâh, Giellagas-instituut

Marja-Liisa Olthuis
Oulu ollâopâttâh, Giellagas-instituut

Fabrizio Brecciaroli
Anarâškielâ servi ry.

Abstract

Inari Saami does not have a strong written tradition. The current orthography was adopted as recently as the 1990s, and the revitalization process is beginning only now to shift its focus from increasing the number of speakers to strengthening the literacy of the language. This article studies the Inari Saami postpositions *pelni* and *piälán* as well as their shorter forms *peln*/*beln* and *pel*/*bel*. The main question is whether these postpositions are joined to the noun preceding them or stand after it as separate words. The research is based on the SIKOR Inari Saami free corpus developed by the Giellatekno team. The postpositions have been analyzed semantically taking into account the frequency with which they occur in the literature. They have been divided into four semantic groups: 1) place, 2) orientation and direction, 3) time and 4) other semantic categories. The long forms *pelni* and *piälán* are mostly written as separate words – except for when they are used to express orientation or direction – whereas the short forms *peln*/*beln* and *pel*/*bel* are usually joined to the preceding word other than in time expressions. Alternative explanations for such variation are also discussed.

Keywords: postposition, orthography, Giellatekno, language revitalisation, language variation

1. Laiđiittâs

Anarâškielâ lii ohtâ Suomâ peln sarnum sämikielâin, mii lii lamaš váduhávt vaarâst lappuđ modern maailm teddui vyelni, mut mii lii lamaš pehtilis iäläskittempargo čuosâttâhhân 1980-lovvoost ovdâskulij. Anarâškielâ iäläskittempargo lii vuáhádum 1900-lovo loopâ peln álgám kielâ renesansân, kuás mielâkiddiivâšvuotâ kielân lasanij (Olthuis 2000: 572–574), kielâpiervâltooimân, mii jođáskij 1990-lovvoost (kj. om. Pasanen 2015) já kielâmiäštártooimân, mii aktivistij puáris kielâsárnoid väldiđ kielâs maasâd kevttimkiellân já sirdeđ tom ovdâskulij L2-sárnoid (Olthuis já iäräseh 2013: 176). Anarâškielâ kevttim lii lasanâm iäläskittem ääigi ennuv já vijđánâm uđđâ areenaid tego tieđâlii čälimân já media kiellân (Olthuis já Trosterud 2015; Olthuis 2021: 3). Onnáá peeivi anarâškielâ iäläskittemtooimâst lii tiäduttum kirjálii kulttuur nanodem ovdâmerkkân kirjálistemprojektijguin, moi ääigi láá uárnejum čälleeškovliittâsah já pyevtittum uđđâ kirjálâšvuotâ (Olthuis já iäräseh 2021: 176–188). Trond Trosterud lii lamaš kuávdáš olmooš taan pargoost kielâteknologin já kielâtotken.

Anarâškielâ čallum kiellân ana ruottâsijdis 1800-lovvoost, mut tääláš ortografia lii kevtiškuottum eskin 1990-lovo aalgâst (Olthuis 2000: 570). Nuorâ čäällimvyevi já -kulttuur keežild čäällimnoormah láá lamaš pääihui ložžâseh, já ovdâmerkkân kuávlukielâliih varianteh já ulmui persovnliih čäällimvyevih láá tiettum čallum käldein kidâ taan peeivi räi. Veikkâ ij liččiigin tárbu stivrid ovtâskâs ulmui kevttim čäällimvyevi meendu äŋgirávt, te čovgâdub noormah láá anoliih ovdâmerkkân oppâmaterialpargoost já virgálijn kielâkevttimsuorgijn tego haldâttâh- já lahâteevstâin, main eksakt olgosadelem lii eromâš tehálâš. Normâdem kuittâg váátá vijđes tiäđu tast, maht kielâ kiävttoo čoođâ kielâsiärváduv – normâdmist kalga väldiđ huámmášumán sehe noormâi já avžuuttâsâi kielâlii systematia (ovdâmerkkân pelikuhes jienâdâh

 Nordlyd 46.1: 171–180, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, toimâttâm Lene Antonsen, Sjur Nørstebø Moshagen já Øystein A. Vangsnes. Almottijjee: UiT Norgga árktalaš universitehta. http://septentrio.uit.no/index.php/nordlyd
https://doi.org/10.7557/12.6384

čalluuččij ain jo-uv uánihâžžân teikkâ kukken) mut meiddei variantij frekvensijd nuuvt et táväliih kielâsiärváduv tuhhiittem sänihäämih iä kuođđuu noormâi ulgguubel. Normâdempargo váátá tiäđu kielâ variaatiost; tuše tubdâmáin kielâ variaatio lii máhđulâš rätkiđ, kalga-uv variaatio siskeldittiđ noormân vâi avžuuttiđ velttiđ tom.

Eromâš tárbu normâdem várás lii šoddâm anarâškielâ kielâteknologisii pargo ooleest. Anarâškielân lii rahtum kielâmalli (Antonsen já iäräseh 2016), mii lii vuáđđun maaŋgâ kielâteknologilii tyejipiergâsân. Táválijd kielâkevtteid já uáppeid kuávdáást láá sänianalysaattorân čonnum sänikirje *Nettidigisäänih* já tivvoomohjelm ađai ortografisâš njuálguluuhâmohjelm. Jis tivvoomohjelm šadda vijđáht kiävtun, te ton merhâšume kielâstivriimân lii uáli styeres: čälleeh veltiškyettih haamijd, maid tot ij tuhhit, já nube tááhust sij uáppih, et tivvoomohjelm tuhhiittem häämihkis láá rievtis häämih. Tivvoomohjelm evaluistmist čielgâi, et tivvoomohjelm lii-uv masa oovtâmielâlâš olmooštivodeijein tast, magarijd haamijd kalga merkkiđ noormâ vuástásâžžân, mut tot lii nube tááhust liijkás čoovgâs já merkkee rievtis haamijd-uv feilân ovdâ-merkkân vájuvái lemmai já vájuváá variaatio myensteristem keežild (Morottaja já iäräseh 2018: 80). Tivvoomohjelm pyeredem váátá tom, et anarâškielâ varijistem kärttejuvvoo já normâduvvoo pyerebeht.

Anarâškielâ kielâmalli keččâluvvoo merikooskâi Kielâtekno-juávhu čokkim anarâškielâ čallum kielâ tekstâčuágálduvváin (SIKOR), mii ana sistees stuorrâ uási puoh čalluin, moh anarâškielân láá almostum uđđâ ortografia ääigi. Taam artikkel čäällimmuddoost kielâmalli máttá analysistiđ 95,59 % tekstâčuágálduv saanijn (rekinistum anarâškielâ analysaattorist 1.12.2021). Analyysijttáá pääccám saanijn uási láá časkem-teikkâ čäällimfeeilah, mut fáárust láá meid ortografia vuástásiih sänihäämih, kielâlâš variaatio teikkâ noormâmiäldásiih sänihäämih, maid kielâmalli lii feilim analysistiđ. Taan nk. *missing list* tarkkuumáin lii máhđulâš karttiiškyettiđ, maggaar fáádá tutkâm ličij puoh ávhálumos kielâmaali pyeredem uáinust.

Ohtâ fáddá, mii missing listist paijaan távjá, lii anarâškielâst tiättojeijee muulsâiävtulâš tääpi čäälliđ tiätu kielâamnâs jo-uv sierâ teikkâ oohtân ovdebáin nominain, om. *peeŋkâ alne ~ peeŋkâaln*, *Aanaar peln ~ Aanaarpeln ~ Aanaarbeln*. Maaŋgah täin kielâamnâsijn láá ärbivuáválávt onnum postposition, mut iä veltihánnáá puoh: Erkki Itkonen lii annaam om. *alne* postposition, mut kielâaamnâs *-pel*/*bel* vist puáhtá leđe jo-uv postpositio teikkâ uási adverbist (ILWB 54, 3191). Suomâ-ugrilii äärbi miäldásávt já meiddei anarâškielâ čäällimkielâ traditiost, mii lii váldám ennuv maali suomâkielâ čäällimnjuolgâdusâin, post-positioh čállojeh tievâdâsâinis sierâ. Kielâamnâs já ovdebáá sääni čäällim oohtân addel kuittâg meiddei máhđulâšvuođâ tagaráid tulkkuumáid, et säänih hämmejeh-uv oovtâst keđgilum adverb, kuálussääni teikkâ joba jieijâs sajehäämi. Tággáár savâstâllâm lii lamaš om. tavekielâ peln tiätu kielâamnâsij peht, tego *-ráigge* (Ylikoski 2014). Suomâkielâ normimpargoost vist lii lamaš savâstâllâm ovdâmeerhâ tiet (-)*päin*-partikkel čälimist sierâ teikâ oohtân, já uđđâsumos miärádâs lii tuhhiittiđ ovddist eenâb oohtânčälimân kyeskee variaatio (kj. Eronen já iäräseh 1996; Kielitoimiston ohjepankki, ohje 129). Oro lemin, et postposition jurdâččum saanij čäällim oohtân tievâdâsâinis sáttá jo-uv kuvviđ ovdeláá mainâšum keđgilumproosees teikkâ tuše ortografisij njuolgâdusâi juárbum. Taan artikkelist mij ep vääldi pele toos, et kuábbáá ääšist lii koččâmuš, mut älkkeevuođâ tiet mij čujottep ovdeláá mainâšum kielâamnâsáid terráin *postpositio*.

Ton čielgim várás, et magareh ärbivuáválávt postposition nobdum säänih läävejeh čalluđ oohtân tievâdâsâinis, mij olášutijm ovdâtutkâmuš já uusâim táválumosijd Morottaja já Olthuis sujâttemoopâst *Inarinsaamen taivutusoppi* (2022) oovdânpuohtum postpositioid tekstâčuágálduvâst sehe sänialgâsávt (postpositio čallum sierâ tievâdâsâinis) já säniloppâsávt (postpositiorááhtus čallum oohtân). Ovdâtutkâmuš vuáđuld mij meridijm valjiđ täärhib tutkâm vuálásâžžân kyehti postpositio variantijdiskuin: *pelni*/*belni ~ peln*/*beln*, *piälán*/*biälán ~ pel*/*bel*/*piäl*. Taah postpositioh láá substantivist *peeli* šoddâm postpositioh/kliitih já toi allegrohámásiih varianteh. Taat juávkku väljejui tutkâmušân nube tááhust toin aggáin, et postpositioin kávnojii ennuv sehe oohtân já sierâ čäällim miäldásiih tábáhtusah, já nube tááhust ton keežild, et rahtum kielâmalli ij piergiittâl tain variaatioin meendugin pyereest já häämih táttuh kiärdášuđ missing listist. Nuuvtpa toh láá kielâteknologisii kielâmaali, kielâtipšom já normâdem tááhust mielâkiddiivááh tutkâmčuosâttuvah. Mij lep kuáđđám tutkâmist meddâl puoh tagarijd *pel*-tábáhtusâid, main lii iäigád koččâmuš lohosääni *peeli* uánánâm häämist já moh tiättojeh tijmeaaigij ohtâvuođâst (om. *tijme lâi* ***pel*** *neelji suulâin*). Siämmáánáál mij lep meddâlistám tábáhtusâid, main *piälán* lii tulkkojum substantiv *peeli* illativhäämmin, om. kuálussaanijn *viljâpiälán*, *tijmepiälán*.

Artikkelstân mij keččâlep västidiđ čuávuváá koččâmušân: Magarijn ohtâvuođâin pajeláá mainâšum säänih tâi säniuásih tiättojeh sierânâs säännin já magarijn ohtâvuođâin toh láá čallum oohtân nominain? Mij

uuccâp vástádâs tutkâmáin postpositioi távjuduv (frekvensijd), tutkâmáin sänihäämi vaiguttâs já olášitmáin ruávis semantlii analyys.

Lovvoost 1 kovvejuvvoo tutkâmvuáđu já -čuolmâ já adeluvvoo tiätu anarâškielâ kielâiäläskittem já kielâtipšom tiileest já táárbust. Lovvoost 2 lää oovdânpuohtum tutkâmamnâstâh já tutkâmmetodeh. Lovvoost 3 mij pyehtip oovdân tábáhtusâid, kuás postpositioh *pelni* já *piälán* variantijdiskuin lää čallum sierâ já oohtân, já mij tutkâp, maht semantlâš analyys čielgee sierâ já oohtân čäällim variaatio. Lovvoost 4 mij kuorâttâllâp analyys.

2. Tutkâmamnâstâh já metodeh

2.1 SIKOR-tekstâčuágáldâh já uuccâmkritereh

Tutkâmamnâstâhân mij kevttip ovdeláá mainâšum Giellatekno-juávhu čokkim anarâškielâ tekstâčuágálduv SIKOR. Taan tutkâmušâst kevttum versio siskeeld 1,77 miljovn säännid. Tekstâčuágáldâh siskeeld iänáážin teevstâid, moh lää lamaš älkkeht finnimnáál digitallii häämist: fáárust lää il. haldâttuvliih já virgáliih teevstah Sämitige siijđoin, oskoldâhteevstah, oppâkirjeh, čaabâkirjáliih kirjeh, motomeh uáppučáittuseh já stuárráámus juávkkun váldu-uási Anarâš-loostâin almostum mainâsijn. Puoh siskeldum teevstah lää čallum 1990-lovo maŋa, kuás uđđâ ortografia valdui anon (SIKOR).

Veikkâ tekstâčuágáldâh lii ärbivuáválii uáinust viehâ ucce, te ucceeblovokielâ uáinust tot kuittâg addel uáli vijđes čáittus taan ääigi anarâškielâst. Mij lep tietimin, et aktiivliih čälleeh iä lah tai peivij räi lamaš nuuvt maaŋgâs (kj. Olthuis já Trosterud 2015), nuuvt et tiätulágán oovtâpiälásâšvuođâ materiaal tulkkuumist kalga väldiđ huámášumán.

2.2 Postpositioi pelni já piälán semantik já ortografisiih konventioh

Inarilappisches Wörterbuch (ađai ILWB, 3191) mield postpositioh *pelni* já *piälán* lää substantivist *peeli* šoddâm postpositioh, adverbeh tâi kliitih, já tain tiättojeh uánánâm varianteh *peln* já *piäl ~ pel*. Mij nomâttep täid uánánâm haamijd allegrohámásâžžân varianttin já haamijd, main loppâvookaal lii siäilum, largohámásâžžân (kj. meid Ijäs 2011).

Taan juávhu saanijn tiättoo meiddei anarâškielân tijpâlâš assimilaationjuolgâdus, mon mield kliiti algâkluusil čálloo čyeijilis jienâduv merháin, jis tot čuávu čyeijilis jienâduv (vrd. lahtospartikkel *-pa/ba*: *moonâmba*, *moonâba*, *moonahpa*), mut taat njuolgâdus ij kuittâg kuoskâ ovdâmerkkân kuálussaanijd (om. *tijmepeeli*, ij **tijmebeeli*)[1]. Nuuvtpa jis postpositioh *pelni* já *piälán* allegrohámásij variantijdiskuin lää čallum oohtân ovdebáin saanijn, te assimilaationjuolgâdus keežild mij sátáččijm pyehtiđ aiccâđ amnâstuvâin meiddei tagarijd haamijd ko *-belni*, *-beln*, *-biälán*, *-biäl* já *-bel*.

Inarilappisches Wörterbuch (ILWB, 3191) addel vuolgâsaje postpositioi *pelni* já *piälán* semantikân, mut eidusiih semantliih kategoriah tai postpositioi várás iä lah anarâškielân huksejum. Hyelkkisääni kategoristem kávnoo pyerebeht suomâkielâ peln[2] já tavekielâst om. Klaus Peter Nickel kielâoopâst *Samisk Grammatikk* postpositioin *bealde* já *beal* (1994: 164, 166, 169) já Pekka Sammallahden *Pohjoissaame-suomi-sanakirja* (Sammallahti: 2020) säniartikkelijn *bealde*, *beal* já *beallái*[3]. Täin käldein pajaneijee táválumoseh semantliih kategoriah, mooid meiddei ILWB:st (ILWB, 3191) kávnojeh ovdâmeerhah, lää 1) saje 2) orientaatio, sunde tâi kuávlu já 3) äigimuddo. Amnâstuv kieđâvušmist mij kiddip huámmášume vuosâsaajeest taaid kategoriaid, mut kejâstep uánihávt meiddei tábáhtussáid, moh iä soovâ tai kategoriai siisâ (juávkku 4).

Tábáhtusâi šlajâttâllâm tai semantlij kategoriai vuálá šadda vuosâsaajeest tarkkuumáin, maggaar lii postpositio tievâdâs. Saje almotteijee postpositioráhtusij tievâdâssân lää ovdâmerkkân enâmij já paihij noomah (om. *Suomâ*, *Aanaar*), merhâšuumeest “kiännii pääihist/pááikán teikkâ kulen/kuuvl“ ulmui

[1]Taan nk. progressivlii assimilaatio lasseen anarâškielâst tiättoo meid regressivlâš assimilaatio, mii lii ucánjáhhá episystematlávt tuhhiittum meid kirjekielân: *čižetpeln ~ čižeppeln*).

[2]kj. om. Kielitoimiston sanakirja (2021), https://www.kielitoimistonsanakirja.fi/#/puoli

[3]Sänikirje nettiversio liiŋkah saanijd: http://satni.org/bealde; http://satni.org/beal; http://satni.org/beallái

jiešnoomah teikkâ ulmuid čujotteijee almosnoomah tâi persovnpronomineh (*Nuuvdi*, *enni*, *mii*) teikkâ meiddei demonstrativ- tâi indefinitpronomineh (*taat*, *kuábáš*, *nubbe*). Orientaatio, sunde já kuávlu almotteijee postpositioráhtusij tievâdâssân láá tijpâlávt nomineh teikkâ adverbeh, moh almotteh tiätulágán relaatioid, ovdâmerkkân *čižet*, *uálgis*, *uulguš*, *siiskiš*, *vuoluš*, *taavaaš*; *teehin*, *tohon*. Äigimudo almotteijee postpositioráhtusij tievâdâssân láá äigimudo almotteijee nomineh tego *kiđđâ*, *ijjâ*, *porgemáánu*, *algâ*, *loppâ* já *1900-loho*.

3. Analyys

3.1 Postpositioi pelni já piälán já toi variantij távjudâh tutkâmamnâstuvâst

Mij uusâim tutkâmamnâstuvâst kuábbáá-uv postpositio *pelni* já *piälán* jyehi variant sehe sänialgâsávt (ađai láá čallum sierâ ovdebáin saanijn)[4] já säniloppâsávt nuuvt et oles sääni ij kuittâg hammii ohtuunis ubâ sänihäämi (ađai lii čallum oohtân ovdebáin saanijn). Mij uusâim meid allegrohámásii postpositio *pel*/*bel* härvinâš variantijd *piäl*/*biäl*, mut toh tiättojii nuuvt harvii (*piäl* kulmii, *biäl* ij ohtiigin) et mij lep ovtâstittám taid tábáhtusâid *pel*/*bel* vuálá. Puátuseh láá puohtum oohtân tavlustuvâst 1.

	Sierâ	**Oohtân**
pelni	166	16
belni	0	0
peln	162	223
beln	0	1102
piälán	46	3
biälán	0	0
pel (piäl)	58	132
bel	1	375

Tavlustâh 1. peeli-vuolgâlij postpositiosanij ovdebáin saanijn sierâ já oohtân čallum tábáhtusâi mereh Korp-amnâstuvâst.

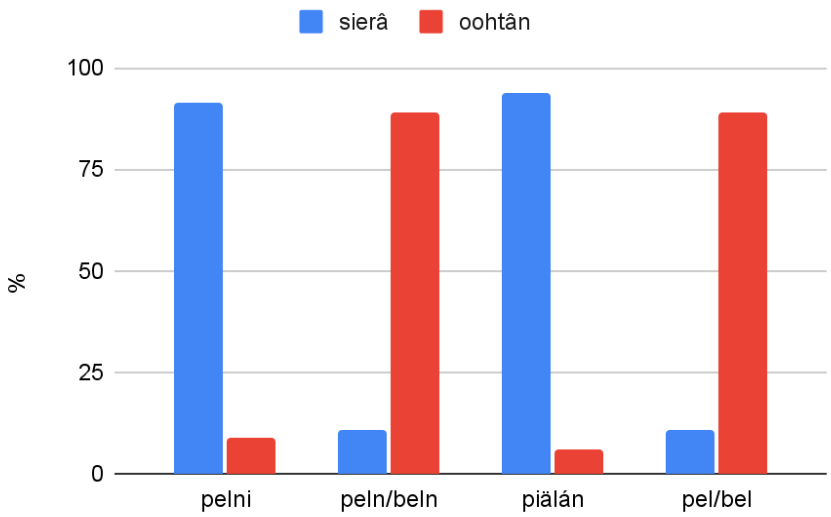
Tavlustuvâst 1 uáiná, et lokativlii postpositio variantijn *pelni ~ belni ~ peln ~ beln* čielgâsávt táválumos lii tievâdâsáin oohtân čallum allegrohámásâš *-beln*. Meiddei illativlii postpositio variantijn (*piälán ~ biälán ~ pel* (*piäl*) *~ bel*) allegrohámásâš *-bel* lii táválumos. Eres varianteh tiättojeh kuittâg meiddei ennuv, eereeb *-belni* já *-biälán* iä tiettuu amnâstuvâst ohtiigin. Čyejihánnáá klusiláin *p* älgee postpositioin tiättoo čielgâ variaatio ton uáinust, et läävejeh-uv toh čalluđ sierâ vâi oohtân tievâdâsâinis.

Vyerdittetteht čyeijilis klusiláin *b* älgee postpositioh vist iä keevâtlávt kuássin tiettuu sierâ säännin tondiet ko tääláš ortografia ij tuárju sänialgâsii *b* (vrd. ILWB kevttim ortografia, 3191)[5]. Sääni siskiibiäläääš *b* vist čuávu anarâškielâ ortografian kullee prinsip tast, et assimilaatio čálloo uáinusân puoh eres sojij peic kuálussaanij loppâuásist.

Kovosist 1 varianteh láá čuákkejum paarrân tađe mield, láá-uv toh largohámásiih vâi allegrohámásiih varianteh, já sierâ já oohtân čäällim variaatio lii oovdânpuohtum prosentuallij koskâvuotâlovoiguin. Largohámásiih *pelni* já *piälán* čállojeh čielgâsávt táválumosávt sierâ ko vist uánánâm *peln*/*beln* já *pel*/*bel* čállojeh iänáážin oohtân – prosentualliih koskâvuotâlovoh láá keevâtlávt largohámásij saanij koskâvuotâlovoi speeijâlkoveh.

[4]CQP-hámásiih uuccâmpákkumeh, ovdâmerkkân *pelni*: sierâ: [word = "pelni"], oohtân: [word = ".*pelni" & word != "pelni"]

[5]Ohtii kavnum sierâ čallum *bel* puáhtá älkkeht tulkkuđ tuše časkemfeilân, veikkâ material kärživuotâ addel-uv ain saje viekkiistâllâđ máhđulâšvuođâ, et tot ličij merkkâ tääláá ortografia juárbumist. Ceelhâohtâvuotâ sierâ čallum *bel*-tábáhtusân lii čuávuvâš: *Juhle maŋa purâdijm já mij (Anssi, Charlie já mun) tolliittijm kuovttijn autoin (Mersuin já Metro-riävoin) neelji ääigi Tave Dakotan já Grand Forks kuávlun já ko tohon peesâim, te tolliittijm Highway 29 mieldi mäádâs, sirdâšuim Maadâ-Dakota bel já puálditttijm 4-rađasii maađij mield ain mäddiláá*.



Kovos 1. peeli-juávhu postpositiosanij sierâ já oohtân čäällim prosentualliih koskâvuotâlovoh.

3.2 Postpositioi oohtân já sierâ čäällim semantlij kategoriai mield

3.2.1 Saje

3.2.1.1 pelni já peln/beln

Amnâstuvâst saje almotteijee *pelni* čálloo ain sierâ tast huolâhánnáá, maggaar täärhib merhâšume ovdebáá säänist lii: om. *Suomâ pelni* (7), *Čovčjäävri pelni* (1), *Aasia pelni* (1), *kaavpug pelni* (1), *Nuuvdi pelni* (1).

Allegrohámásii postpositiost *peln* tiättojeh sehe sierâ já oohtân čallum tábáhtusah, mut oohtân čäällim lii ennuv táválub. Ovdâmerkkân talle ko ovdebâš substantiv lii eennâm nommâ, te amnâstuvâst lii kavnâmist sehe sierâ čallum *Suomâ peln* (1) mut meid oohtân čallum *Suomâpeln* (12). Allegrohámásijd oohtân čallum tábáhtussáid puáhtá luuhâđ meid *beln*-tábáhtusâid, moh láá valjeest, já ovdâmerkkân *Suomâbeln* tiättoo-uv amnâstuvâst 34 kerd. Nuuvtpa tievâdâs *Suomâ* puotâ sierâ čallum tábáhtusah láá allegrohámásijn saanijn ohtsis 1 já oohtân čallum tábáhtusah láá 46. Sierâ čallum tábáhtus ij oro lemin tuše ovtâskâs časkemfeilâ, tondiet ko saje almotteijee *peln*-tábáhtusah kávnojeh kale sierâ čallum eres sanij puotâ, nuuvt ko *Taažâ peln* (1), *Aanaar peln* (1), *Ucjuv peln* (1).

Kiännii pääihi almotteijee postpositioh iä čuávu aaibâs siämmáá tendens: amnâstuvâst tiättoo stuárráb variaatio ton uáinust, et čálloo-uv allegrohámásâš postpositio sierâ vâi oohtân: om. *Nuuvdi peln* (3) – *Nuuvdibeln* (12), mut *mii*/*muu peln* (8) – *miibeln* (9), *eeni*/*iännán*/*eenis peln* (5) – *eenipeln*/*eenibeln* (0).

Ennuv variaatio lii meid allegrohámásijn postpositioin talle, jis ton tievâdâssân lii genetivhäämi *nube*: sierâ čallum *nube peln* tiättoo amnâstuvâst 39 kerdid já oohtân čallum *nubepeln* ~ *nubebeln* tiättoo 49 kerdid.

Saje almotteijee lokativlij *pelni* já *peln*/*beln* tievâdâsâi täärhib tarkkum čáittá, et largohámásâš lokativlâš postpositio läävee čalluđ sierâ almolii tendens mield, mut allegrohámásijn postpositioin lii stuorrâ iäru variaatiost ton mield, maggaar sääni lii postpositio tievâdâssân.

3.2.1.2 piälán já pel/bel

Largohámásâš *piälán* ij tiettuu amnâstuvâst meendu távjá, mut toh tábáhtusah moh tiättojeh saje almotteijee merhâšuumeest, láá puoh čallum sierâ, om. *Paččvei piälán* (2), *Syyria piälán* (1), *kuábbáá-uv piälán* (1), *nube piälán* (1), *taan piälán* (1), *tii piälán* (1).

Allegrohámásii postpositiost *pel* tiättojeh oohtân čallum tábáhtusah mottoomverd eenâb ko sierâ čallum tábáhtusah, mut variaatio lii siämmáásullâsâš tievâdâsâst sorjohánáá: *Suomâ pel* (3) – *Suomâbel* (4), *Taažâ pel* (6) – *Taažâbel* (8), *nube pel* (13) – *nubebel*/*nubepel* (51), *mii pel* (4) – *miibel* (12).

Saje almotteijee illativlij *piälán* já *pel*/*bel* tievâdâsâi täärhib tarkkum čáittá, et largohámásâš illativlâš postpositio läävee čalluđ sierâ almolii tendens mield, mut allegrohámásijn postpositioin lii ennuv variaatio, mii ij kuittâg sorjo tast, maggaar sääni lii postpositio tievâdâssân. Taat aiccâmuš oro tohâmin iäru lokativlii *peln* já illativlii *pel* kooskâst.

3.2.2 Orientaatio, sunde já kuávlu

3.2.2.1 pelni já peln/beln

Taan semantlii kategorian kullee postpositio *pelni* epivyerdittetteht ij tiettuugin amnâstuvâst táválávt sierâ čallum häämist, pic tot čálloo täävjib oohtân. Ovdâmerkkân tábáhtus *ulg(g)uu pelni* tiättoo tuše ohtii, já *ulg(g)uupelni* vist tiättoo neljii; áimusuundijn sierâ čallum tiättojeh *taveuárji pelni* (1) já *tavenuorttii pelni* (1), mut oohtân čallum tábáhtusah-uv kávnojeh kyehti: *taavaapelni* (1) já *nuorttiipelni* (1). Ton lasseen oohtân čallum tábáhtusah láá *uálgispelni* (1), *vuoluubelni* (1), *čižetpelni* (1) já *tuárispelni* (1). Merhâšittee lii kale mainâšiđ, et largohámásiih *pelni*-tábáhtusah ubânâssân iä lah meendu ennuv amnâstuvâst, já sunde almotteijee postpositio amnâstuvâst láá-uv täävjib čallum allegrohámásijn postpositioin *peln*/*beln*.

Tego vuordum, allegrohámásiih postpositioh čállojeh täävjib oohtân ko sierâ. Almolâš aiccâmuš lii, et tiätulágán saanij ohtâvuođâst variaatio lii uccáá teikkâ ij ollágin: ovdâmerkkân oohtân čallum *ulg(g)uupeln* ~ *ulg(g)uubeln* tiättojeh ohtsis 274 kerdid, já *ulg(g)uu*-saanijn sierâ čallum tábáhtusah iä kavnum ohtâgin; *čižetpeln* ~ *čižetbeln* tiättojeh ohtsis 37 kerdid já sierâ čallum *čižet peln* tiättoo ohtii; *seelgi peln* tiättoo ohtii já *seelgipeln* ~ *seelgibeln* tiättoo ohtsis 13 kerdid. Taan sullâsijn saanijn oroh lemin vuáhádum čäällimkonventioh.

Taan semantlii kategorian kullee postpositioi sierâ já oohtân čäällim iäráneh almolii tendensist tienuuvt, et largohámásâš čálloo-uv täävjib oohtân siämmáánáál ko allegrohámásiih postpositioh-uv. Ton lasseen allegrohámásijn postpositioin tiettui stuorrâ iäru variaatiost ton kuáttá, et mii saanijd postpositio oovdeld lii.

3.2.2.2 piälán já pel/bel

Largohámásâš *piälán* kiävttoo taan semantlii kategoriast tuše marginallávt, mut puoh kavnum tábáhtusah čuávuh lokativlii *pelni* maali: oohtân láá čallum *ovdiipiälán* (1), *ulguupiälán* (1) já *čižetpiälán* (1). Sierâ čallum tábáhtusah iä tiettuu amnâstuvâst. Siämmáš kuáská meid allegrohámásii postposition *pel*: oohtân čallum tábáhtusah láá om. *ulg(g)uupel*/*-bel* (70), *čižetpel* (18) já *taavaapel*/*taavaabel* (22). Sierâ čallum tábáhtusah iä tiettuu amnâstuvâst.

Taan semantlii kategorian kullee illativlijn postposition ij lah mainâšittee variaatio ton uáinust, et čállojeh-uv toh oohtân vâi sierâ tievâdâsâinis.

3.2.3 Äigi

3.2.3.1 pelni já peln/beln

Ääigi almotteijee largohámásâš *pelni* čálloo amnâstuvâst tiätu tievâdâsâiguin ain sierâ: om. *1800-lovo pelni*, *1900-lovo pelni*, *2000-lovo pelni* já *1980-lovo pelni* (ohtsis 12), *porgemáánu pelni*, *kesimáánu pelni* (ohtsis 2), *čoovčâ pelni* (1), *kiiđâ pelni* (1). Smavvâ variaatio tiättoo abstraktlub äigimudo kovvejeijee tievâdâsâi *algâ* já *loppâ* puotâ: oohtân čallum tiettui *aalgâpelni* ~ *algâpelni* (2), já *loopâpelni* (2). Sierâ čallum *aalgâ pelni* ij tiettum ohtiigin, *loopâ pelni* vittii.

Ääigi almotteijee allegrohäämi čálloo meid táválávt sierâ: *syeinimáánu peln*, *kuovâmáánu peln*, *roovvâdmáánu peln* (6), (*moonnâm/taan/puáttee*) *ive peln* (8), *kiiđâ peln* (1). Variaatio tiättoo kuittâg eenâb ko largohámásii postpositiost, ovdâmerkkân sierâ lii čallum *eehid peln* (1) mut meiddei oohtân *eehidbeln* (2). Siämmáánáál ko largohámásii postpostio peht, tievâdâsah *algâ* já *loppâ* spiehâsteh taan tendensist:

sierâ čallum *aalgâ peln* ij tiettuu ohtiigin, *loopâ peln* tuše ohtii, mut *aalgâbeln* tiättoo 26 kerdid já *loopâbeln ~ loppâbeln* 35 kerdid.

Ääigi almotteijee lokativlij *pelni* já *peln*/*beln* tarkkum čáittá, et sehe largohámásâš já allegrohámásâš lokativlâš postpositio läävee čalluđ sierâ já máhđulii variaatio stivree vuosâsaajeest tievâdâs.

3.2.3.2 piälán já pel/bel

Sehe largohámásâš *piälán* já allegrohámásâš *pel*/*bel* kiävttojeh ääigi almotteijee merhâšuumeest uáli harvii. Largohámásâš *piälán* lii kyevti tábáhtusâst čallum sierâ já ij ohtiigin oohtân: *eehid piälán* (1), *kiđâ piälán* (1). Allegrohámásâš *pel* vist tiättoo tuše ohtii, já ton tábáhtusâst tot lii čallum oohtân: *loopâpel* (1). Taat puáđus tuárju jieijâs uásild jurduu, et allegrohámásiih postpositioh čállojeh täävjib oohtân ko largohámásiih, mut tábáhtusâi vyeligis mere keežild puáđus lii tuše marginallâš.

3.2.4 Eres semantliih kategoriah

Amnâstuvâst tiättojeh mottoomverd postpositioi *pelni* já *piälán* tábáhtusah, moh iä merhâšuumees táához kuulâ kuulmâ váldukiävtu (saje, orientaatio já äigi) siisâ. Tagarij tábáhtusâi meeri amnâstuvâst lii viehâ ucce, já tijpâlávt tiätu tievâdâsâi teikkâ semantlávt oohtân čonnâšum tievâdâsjuávhui ohtâvuođâst kevttum postpositioin iä kavnuu sehe largo- já allegrohámásiih tábáhtusah kuábbáá-uv postpositio várás. Nuuvtpa toi täärhib semantlâš šlajâttâllâm ij puávtáčči ennuvgin lasetiäđu sierâ já oohtân čäällim variaatiost. Taan lovvoost mij oovdânpyehtip uánihávt kyehti kieđâvušhánnáá kuođđum semantlii kategoria.

Vuossâmuš juávkku lii tiätulágán abstraktlâš aldaašvuođâ teikkâ oohtânkulleevâšvuođâ almotteijee merhâšume, mii lii sáttám vuáháduđ meid idiomatlâžžân. Tágáreh tábáhtusah amnâstuvâst láá om. *vuáitu pelni*, *puollâš pelni* já *väldidem pelni*.

Nubbe juávkku kovvee mottoomlágán juávkun kuullâm teikkâ juávhu tuárjum. Tágáreh tábáhtusah láá om. *jurduu pelni* (*Veikkâ Takkusiist lâi-uv taggaar ruopsis pottááknjune, te sun-uv lâi ton vijnettes já tubbáákttes jurduu pelni, veikkâ njune ivne ličij maid peri tuođâštâm.*), *mii pelni* (*Jis Immeel lii mii pelni, kii puáhtá leđe mii vuástá?*), *eeji pelni* (*sun tobdá jieijâs rievtis sämmilâžžân, veik sun lii sämmilâš tuše eejis pelni.*) já *Immeel piälán* (*Puoh vädisvuođâin mij kolgâp jurgâluđ Immeel piälán já vyerdiđ suu iše.*) Taam juávhu tubdâstem lii tehálâš, tondiet ko tievâdâsâidis táához tot sulâstit saje kategoria tábáhtusâid merhâšuumeest “kiännii pääihist/pááikán t. kulen/kuuvl” já nuuvtpa oles ceelhâohtâvuođâ tarkkum lii lamaš tehálâš. Motomin tábáhtusah sättih liijká leđe semantlii kategoria táához ambivalenteh, veikkâ ceelhâohtâvuotâ ličij-uv tiäđust.

4. Kuorâttâllâm

Toi tábáhtusâi frekvensij rekinistem, kuás postpositioh *pelni ~ peln ~ beln* já *piälán ~ pel ~ bel* láá čallum tievâdâsâinis sierâ já kuás oohtân, puáhtá uáinusân čielgâ tendens tast, et largohámásiih postpositioh čállojeh masa ain sierâ já allegrohámásiih postpositioh čállojeh masa ain oohtân. Tom čuávvum semantlâš tarkkum čáittá ton lasseen čuávuváá:

a. Jis koččâmušâst lii saje almotteijee postpositiorááhtus, te allegrohámásijn postpositioin lâi eenâb variaatio (mut largohámásijn ij lah), já lokativlijn postpositioin variaation vaaigut ennuv tot, maggaar sääni lii postpositio tievâdâssân (mut illativlijn ij).

b. Orientaatio, sunde tâi kuávlu almotteijee postpositioráhtuseh spiehâsteh almolii tendensist nuuvt, et sehe largo- já allegrohámásiih postpositioh čállojeh masa ain oohtân. Meid taan juávhust allegrohámásijn postpositioin lii meid eenâb variaatio ko largohámásijn, mut suijân puáhtá leđe meid tot, et largohámásiih postpositioh tiättojeh mudoi-uv ucceeb ubâ amnâstuvâst.

c. Ääigi almotteijee postpositioráhtusijn sehe largo- já allegrohámásiih postpositioh čállojeh tijpâlávt sierâ já máhđulii variaatio stivree vuosâsaajeest kevttum tievâdâs. Stuárráámus spiehâstâh tiättoo eromâšávt kyevti tievâdâs *algâ* já *loppâ* peht, moh čállojeh masa ain oohtân.

Pajeláá mainâšum puátusijd puáhtá čoonnâđ oohtân tienuuvt, et sehe sääni häämi (largo-allegro) já postpositioráhtus merhâšume (saje-orientaatio-äigi) čielgejeh stuorrâ uási tast, láá-uv postpositioh čallum sierâ vâi oohtân tievâdâsâidiskuin, mut meiddei ovtâskâs tievâdâsâin lii merhâšume.

Puátusij vuáđuld lii máhđulâš kuorâttâllâđ koččâmuš, et láá-uv taan artikkelist postposition nobdum kielâamnâseh ärbivuávälii jurdâččemvyevi mield postpositioh, kuálussäänih, keđgilum adverbeh vâi uđđâ, moráneijee sajehäämih. Lii máhđulâš, et toh tuáimih postposition tiätulágán merhâšuumeest (om. äigi), mut mottoom nube merhâšuumeest (orientaatio) postpositioráhtuseh láá hammim keđgilum adverb. Lii meiddei máhđulâš, et allegrohámásii postpositio tiettum táválubbooht oohtân čälimist lii iäigád keđgilumproosees čuávumuš, já tekstâčuágálduvâst tiättojeijee teevstâi čälleeh láá valjim čäälliđ saanijd oohtân talle ko sij aneh taid adverbijd tijpâlii vyevi miäldásávt oohtân čonnâšum, vuáhádum já toppum sänijuávkkun.

Oohtân čäällim puáhtá leđe merkkâ tast, et čällee ana ráhtus kuálussäännin teikkâ tot lii onnum kuálussäännin ovdil ko tot finnij postposition tijpâlijd jiešvuođâid. Tondiet ko tutkum postpositioh láá vuálgus substantivist *peeli*, mii tiättoo ennuv kuálussanij váldu-uássin, lii máhđulâš tulkkuđ nuuvt, et maaŋgâ mii tutkâmuš tábáhtusâst vuáđđun lii-uv riävtui kuálussääni, om *algâbeln* < *algâpeeli*, *loppâbeln* < *loppâpeeli*, *čižetpeln* < *čižetpeeli*, *uálgispeln* < *uálgispeeli*. Eromâšávt almolii tendensist spiehâsteijee tábáhtussáid tot adeličij hiäivulii čielgiittâs. Nubbe tulkkumvyehi ličij tot, et tágáreh sääniih láá tuše litodâttâm čovgâsávt merhâšuumees tááhust oohtân, nuuvt et taid lii älkkee anneeđ kuálussäännin (kj. meid Kielitoimiston ohjepankki, ohje 112). Tágárijn saanijn lii jo-uv nominativmiärus, tego saanijn *algâpeeli* já *loppâpeeli*, teikâ genetiivmiärus tego *ulg(g)uubeln* (*uulguš*: gen. *ulg(g)uu*). Kuálmád čielgiittâs oohtân-čälimân sáttá leđe oovtâkiärdánávt tuše tot, et mon táválâš tiätu sääni siskeldeijee postpositiorááhtus lii: táválub sänilitoh suddeh älkkeebeht oohtân já leksikalisistojeh.

Tot et saje almotteijee tábáhtusâin ij tiettuu siämmáá čielgâ sunde ij sierâ ige oohtân čälimân ko kyevti eres kategoriast, sátáččij merhâšiđ jo-uv tom, et keđgilumproosees lii ton kategoria peht eskin joođoost, teikkâ tom, et semantlâš kategoria lii nuuvt viijđes, et toho čäähih vyelijuávhuh, moi kooskâ jiešalnees liččii-uv tiettum čielgâ iäruh. Ohtâ iävtukkâs täärhib semantlii juávu merhâšuumeest itá saje almotteijee vyelijuávhust "kiännii pääihist/pááikán tâi kiännii kulen/kuuvl". Taan juávhust sierâ lii ovdâ-merkkân čallum *Nuuvdi peln* já *mii peln*, mut siämmái tievâdâsâiguin čállojeh meid oohtân *Nuuvdibeln*, *Nuuvdibel* já *miibeln*. Jieččân kielâfiätust paijaan potentiallâžžân čielgiittâssân tot, et sierâ čallum tábáhtusâin *Nuuvdi peln* já *mii peln* máhđulávt referenttin lii olmooš (*Nuuvdi*, *mij*) já vist tábáhtus *Nuuvdibeln* oroččij pyerebeht-uv čuujootmin táálun (sajan) já ton peht táálust ässee ulmui(d).

Tast huolâhánnáá, et sänihäämi já semantlâš šlajâttâllâm pyevtitteh čielgâ puáttus, te tijpâlâš lii kuittâg, et jyehi tutkum juávhust lii variaatio, mon taah kyehti jiešvuođâ iä čielgii. Taggaar variaatio puáhtá anneeđ nuorâ čäällimkulttuur já siemin siärváduv čuávumuššân. Vistig-uv variaatio puáhtá čielgiđ oovtâkiärdánis analogia: jis postpositio lii monnii tiätu uápis ohtâvuođâst čallum távjá oohtân, te tot sáttá čalluđ oohtân meiddei mottoom nube kiävtust, veikkâ saanijd ij anaččiigin siämmáá čielgâsávt ohtân-kulleevâžžân. Puáhtá vävjiđ, et tággáár kontaminaatio tiättoo eenâb hárjánmettumis čällein já L2-čällein, kiäh iä lah vanttâm luuhâđ iäge tuubdâ saanij semantik siämmáá pyereest ko kielâmiäštáreh. Viereskielâliih L2-čälleeh láá kuittâg váldám aktivlii rooli anarâškielân čälimist. Eenikielâliih čälleeh láá tuše váájuv lovmat ubâ kielâsiärvusist (Olthuis já Trosterud 2015; Olthuis já iäräseh 2021).

Tutkâmamnâstuvâst šaddee vyerdimettumis variaatio teikkâ tendensijd – ovdâmerkkân ovtâskâs tievâdâsâi spiehâstem almolijn tendensijn – puáhtá čielgiđ tot, kiäh láá lamaš teevstâi čälleeh. Anarâškielân čälleeh láá uccáá já lii tiäđust, et tekstâčuágálduv teevstâin stuorrâ uási láá vuálgus tuše oovtâ ulmust teikkâ láá jottáám kielâtipšom čoođâ, mast lii iänáážin västidâm nubbe. Ovtâskâs ulmuu idiolektist oohtân já sierâ čälimân vissásávt vaaigut tot, mon ennuv sun uáiná teikkâ kiävttá postpositioráhtus já mon tehálâš äšši tot čällee persovnlii maailmist lii. Taam tekstâčuágálduv jiešvuođâ puáhtá anneeđ tiätulágán hiäjuvuottân já pajediđ koččâmuš, et puáhtá-uv tekstâčuágálduv puátusijd anneeđ šiev váldusin eidusii anarâškielâst. Mij halijdep kuittâg tiäduttiđ, et tekstâčuágálduv kielâ ij lah tuše muádi ulmuu kielâ, mut lii vijđáht tuhhiittum kielâsiärváduvâst rievtis já riges anarâškiellân.

5. Loopân

Mii tutkâmuš oonij sistees tuše kyehti postpositio, mon mij valjijm vuosâsajasâžžân tutkâmčuosâttâhhân ovdâtutkâmuš vuáđuld. Siämmáá ovdâtutkâmušâst pajanii eenâb-uv šiev tutkâmčuosâttuvah, main tiettui oohtân já sierâ čälimân kyeskee variaatio. Mij ferttijm vaidâlitteht raijiđ taid olgos taan tutkâmušâst, mut puátteevuođâst ličij ávhálâš vijđediđ postposition nobdum largo- já allegrohámásij ráhtusij oohtân já sierâ čäällim tutkâm ovdâmerkkân postpositioperruid *alne* ~ *alne*, *oolâ* ~ *ool*, *pehti*/*behti* ~ *peht*/*beht*, *náálá* ~ *náál*, *kulen*/*gulen* ~ *kuvlân*/*guvlân* já *kuuvl*/*guuvl*. Ovdâtutkâmušâst postpositioperrust *siisâ* ~ *siis*, *siste* ~ *sist* vist tiettui mielâkiddiivâš ovdâmerkkâ oohtân teikkâ sierâ čälimist: sehe largohámásiih já allegrohámásiih postpositioh lijjii ain čallum sierâ, peic *siisâ* lâi kuittâg čallum oohtân talle ko lâi koččâmuš peljimeerhâin (om. *rastasiisâ*, *skiivŋesiisâ*). Taat lii šiev ovdâmerkkâ adverbin keđgilum postpositioráhtusist.

Ohtâ tutkâmuš vuolgâsoojijn lâi tárbu lasettiđ tiäđu anarâškielâ variaatiost. Čoggâšum tiäđu lii tehálâš heiviittiđ ovdâmerkkân kielâ normâdmist já kielâtipšomist, amas noormah šoddâđ iänáážin oovtâ ulmuu, kuávlukielâ teikkâ juávhu vuáđuld. Taan tutkâmuš juurdâpuáđus normâdem vuáđđun puáhtá leđe, et táváliih tábáhtusah já almoliih tendenseh annojeh vuosâsajasâžžân häämmin – njuolgâdussân om. puoh largohámásiih postpositioh čállojeh tievâdâsâidiskuin sierâ já puoh allegrohámásiih oohtân, teikkâ ääigi almotteijee postpositioráhtuseh čállojeh sierâ já orientaatio tâi sunde almotteijee postpositioráhtuseh čállojeh oohtân. Nube tááhust kuittâg variaatio tiettum lii siämmást argument ton peeleest, et távjá ij lah tuše ohtâ häämi, mon kielâkevtteeh aneh olmâ kiellân. Variaatiotutkâmuš juurdâpuátusin puáhtá leđe meiddei tuhhiittiđ variaatioid vijđáht já faallâđ tiäđu kielâkevttim máhđulâšvuođâin já tast, magarijd nyansijd oohtân já sierâ čäällim sátáččii anneeđ sistees.

Ohtâ artikkel čällein (Olthuis 2008) čáálá nube ohtâvuođâst tast, mon herkis tiileest anarâškielâ iäläskittem lii lamaš já tast, et korrâ kielâtipšom já purism älkkeht tipšoh kielâ toppâlum tilán. Lii jiärmálub tuhhiittiđ valjab variaatio ko leđe ollásávt čälihánnáá ucceeblovokielân.

Käldeeh

Antonsen, Lene, Trond Trosterud, Marja-Liisa Olthuis já Erika Sarivaara. 2016. Modelling the Inari Saami morphophonology as a finite state transducer. Almostitmist *The Second International Workshop on Computational Linguistics for Uralic Languages:* 3–13, toimâttâm Tommi A. Pirinen, Eszter Simon já Francis M. Tyers já Veronica Vincze. Szegedi Tudományegyetem, Szeged. https://www.academia.edu/72593469/Report_on_the_Second_International_Workshop_on_Computational_Linguistics_for_Uralic_Languages

Eronen, Riitta, Sari Maamies já Anneli Räikkälä. 1996. Yhdyssanat. *Kielikello* 4/1996. https://www.kielikello.fi/-/yhdyssanat. [Čujottum 28.3.2022.]

Ijäs, Johanna Johansen. 2011. *Davvisámegiela finihtta vearbahámiid sojahanvuogádaga oččodeapmi vuollel golmmajahkásaš máná gielas*. Davvi Girji, Kárašjohka.

ILWB = Itkonen, Erkki (hrsg.), Raija Bartens já Lea Laitinen (unter Mitarbeit). 1986–1991. *Inarilappisches Wörterbuch* I–IV. Lexica Societatis Fenno-Ugricae XX. Suomalais-Ugrilainen Seura, Helsinki.

Kielitoimiston ohjepankki: *Yhdyssana vai ei*. http://www.kielitoimistonohjepankki.fi/ohje/112. [Čujottum 9.12.2021.]

Kielitoimiston ohjepankki: *Yhdyssana vai ei*. http://www.kielitoimistonohjepankki.fi/ohje/129. [Čujottum 9.12.2021.]

Kielitoimiston sanakirja. 2021. Kotimaisten kielten keskuksen verkkojulkaisuja 35. Kotimaisten kielten keskus, Helsinki. URN:NBN:fi:kotus-201433. https://www.kielitoimistonsanakirja.fi. Peividuvvee almosittem. Peividum 11.11.2021 [Čujottum 08.12.2021.]

Morottaja, Petter, Marja-Liisa Olthuis, Trond Trosterud já Lene Antonsen. 2018. Anarâškielâ tivvoomohjelm. Kielâ-já ortografiafeeilâi kuorrâm tivvoomohjelmáin. Almostitmist *Gielladieđalaš čállin sámegillii – Gii das berošta? Dutkansearvvi dieđalaš áigečála* Vol. 2 Issue 2: 63–84, toimâttâm Marja-Liisa Olthuis já Irja Seurujärvi-Kari. https://www.dutkansearvi.fi/volume-2-issue-2-fi/.

Morottaja, Petter já Olthuis, Marja-Liisa 2022: *Inarinsaamen taivutusoppi*. Sämitigge, Aanaar.

Nettidigisäänih. Anarâškielâ sänikirje. 2021. UiT The Arctic University of Norway, Giellatekno, Tromsø. Peividuvvee almostittem. https://saanih.oahpa.no/ [Čujottum 08.12.2021.]

Nickel, Klaus Peter. 1994. *Samisk grammatikk*. Davvi Girji, Kárašjohka.

Olthuis, Marja-Liisa. 2000. Inarinsaamen kielen vuosisadat. *Virittäjä*, Vol. 104 nr. 4: 568–575.

Olthuis, Marja-Liisa. 2008. Inarinsaamen huoltoa ja elvytystä. *Kielikello* 1/2008. https://www.kielikello.fi/-/inarinsaamen-huoltoa-ja-elvytysta.

Olthuis, Marja-Liisa, Suvi Kivelä já Tove Skutnabb-Kangas. 2013. *Revitalising Indigenous Languages – How to Recreate a Lost Generation*. Multilingual Matters, Bristol. https://doi.org/10.21832/9781847698896.

Olthuis, Marja-Liisa já Trond Trosterud. 2015. Inarinsaamen lingvistinen suunnittelu kieliteknologian valossa. Almostitmist *Vähemmistökielten revitalisaatio*. AGON n:o 45–46. 1–2/2015, toimâttâm Annika Pasanen já Sanna Valkonen. http://agon.fi/article/inarinsaamen-lingvistinen-suunnittelu-kieliteknologian-valossa/.

Olthuis, Marja-Liisa, Trond Trosterud, Erika Katjaana Sarivaara, Petter Morottaja já Eljas Niskanen. 2021. Strengthening the Literacy of an Indigenous Language Community: Methodological Implications of the Project Čyeti čälled anarâškielân 'One Hundred Writers for Aanaar Saami'. Almostitmist *Indigenous Research Methodologies in Sámi and Global Contexts. New Research, New Voices*: 175–200, toimâttâm Pirjo Kristiina Virtanen, Pigga Keskitalo já Torjer Olsen. Brill Sense, Leiden. https://doi.org/10.1163/9789004463097_008.

Olthuis, Marja-Liisa. 2021. Nubástusâi čohčâ. *Anarâš* 2/2021:3. https://drive.google.com/drive/folders/0ByPNya2iI49hcGI1RXZ2MV9LbFU?resourcekey=0-1Su9TqVEoPYNlgiHzkxWVQ.

Pasanen, Annika. 2015. *Kuávsui já peeivičuovâ. 'Sarastus ja päivänvalo' : Inarinsaamen kielen revitalisaatio.* Uralica Helsingiensia 9. Suomalais-Ugrilainen Seura ja Helsingin yliopisto, Helsinki.

Sammallahti, Pekka. 2020. *Pekka Sammallahden pohjoissaame-suomi-sanakirja*. Divvun, Romsa. [Čujottum 9.12.2021.] http://satni.org/sammallahtismefin.

SIKOR. UiT Norgga árktalaš universitehta ja Norgga Sámedikki sámi teakstačoakkáldat, Veršuvdna 01.10.2021, http://gtweb.uit.no/korp/.

Ylikoski, Jussi. 2014. Davvisámegiela *-ráigge* – substantiiva, advearba, postposišuvdna vai kásus? *Sámi Dieđalaš Áigečála* 2/2014: 47–70. https://site.uit.no/aigecala/files/2015/04/SDA-2-2014-ylikoski.pdf.

Språkdokumentasjon innen fennistikken og kvensk

Leena Niiranen
UiT Norges arktiske universitet

Abstract

The study of the Finnish language – called Fennistics – focused on collecting Finnish dialect material from very early on. During the 19th century the interest in studying dialects was governed by the idea that dialects could be used to develop modern written Finnish. However, gradually the study of dialects also became an area of study in its own right. Collecting material on Kven dialects belonged to the larger project of Fennistic data collection from the very beginning. As a consequence of this, a substantial amount of material about Kven dialects can be found in Finnish dialect archives, material which has been used in the process of revitalizing Kven. In this article, language documentation is defined as an activity which also includes traditional dialectology. By contrast, documentary linguistics is a field of linguistics established in the 1990s which focuses on revitalizing endangered languages. The paper compares and discusses these two approaches to language documentation.

Keywords: Kven, dialectology, language documentation, Fennistics

1. Innledning

Språkdokumentasjon er en viktig del av språkrevitalisering. Spesielt i situasjoner der språket ikke lenger brukes aktivt, må for eksempel utviklingen av læremateriell basere seg på tidligere språkdokumentasjon i arkiver. I denne artikkelen diskuterer jeg betydningen av språkdokumentasjon av kvensk i finske arkiver for revitaliseringen av kvensk språk.

Trond Trosterud har gitt et viktig bidrag til kvensk språkdokumentasjon gjennom sitt arbeid med kvensk språkteknologi. Arbeidet hans inkluderer etableringen av elektronisk kvensk ordbok og en morfologisk analysator for kvensk (Giella-fkv). Dette er nyttige hjelpemidler for alle som ønsker å lære seg kvensk, og ikke minst for studenter som studerer kvensk ved UiT (Giellatekno).

Selv om dokumentasjon av kvensk også har skjedd i Norge, både før og etter at kvensk fikk status som eget språk i 2005, vil jeg i denne artikkelen konsentrere meg om dokumentasjonen av kvenske dialekter som finnes i finske arkiver. Mye materiale av kvenske dialekter i finske arkiver skyldes at de ble ansett for å være dialekter av finsk før kvensk fikk status som eget språk (Hyltenstam og Milani 2003; Söderholm 2017: 23).

Først vil jeg presentere begrepet, *språkdokumentasjon*, og hvordan dette begrepet skiller seg fra *dokumentasjonslingvistikk.*[1] Jeg drøfter også betydningen av språkdokumentasjon for språkrevitalisering. Påfølgende presenterer jeg hvordan kvensk språk har inngått som en del av prosjekter for innsamling av materiale innen fennistikken, altså forskningen på det finske språket. Målet er å gi en oversikt over ulike samlinger av kvenske dialekter i finske arkiver. Jeg vil drøfte hvordan og med hvilken hensikt disse materialene ble samlet inn. Avslutningsvis drøfter jeg betydningen av denne språkdokumentasjonen for revitaliseringen av kvensk språk.

[1] På engelsk *language documentation* og *documentary linguistics*.

 Nordlyd 46.1: 181–192, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigert av Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publisert ved UiT Norges arktiske universitet. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6347

2. Dokumentasjonslingvistikk, språkdokumentasjon og revitalisering

På 1990-tallet økte oppmerksomheten på at mange språk var truet av å forsvinne ut av bruk. Austin og Sallabank (2018: 207) konstaterer at språkdokumentasjon ble etablert blant lingvister som en reaksjon på denne situasjonen. Lingvistikkens oppgave var å dokumentere truede språk. I tillegg ble revitalisering av truede språk også en agenda for mange språkforskere.

Målet for språkdokumentasjon av truede språk er å samle inn et representativt korpus av språk brukt i dets sosiale og kulturelle kontekst. Målet er også at korpuset skal reflektere variasjon mellom språkbrukere i ulike alder og kjønn. Talere av truede språk skal også delta aktivt i innsamlingsprosjekter i tillegg til forskere (Austin 2020: 199–209).

Mens Austin og Sallabank (2018) og Austin (2020) betrakter språkdokumentasjon som et nytt fagfelt, er språkdokumentasjon etter Woodbury (2011: 159) like gammel virksomhet som skrivekunsten. Hans definisjon av språkdokumentasjon er «opprettelse, annotasjon, bevaring og formidling av transparente opptegnelser av et språk»[2]. Austin (2020: 203–205) derimot kaller resultater av tidligere innsamlingsprosjekter for 'eldre / tradisjonsmaterialer' «legacy materials». Selv om dette materialet enkelte ganger er samlet inn av profesjonelle lingvister, kan deres kvalitet variere. Slikt materiale kan likevel være verdifulle for revitaliseringsprosjekter i tilfeller det ikke lenger finnes språkbrukere igjen.

Jalava og Sandman (2018: 597–598) påpeker at lingvistisk feltarbeid innen amerikansk strukturalisme var det historiske utgangspunktet for dokumentasjonslingvistikk, og at feltarbeid i tillegg var vanlig innen fennougristikk. Woodbury (2011: 162–163) løfter frem spesielt den amerikanske antropologien Franz Boas som allerede på begynnelsen av 1900-tallet utviklet metodikk for språkdokumentasjon, og derfor kan betraktes som en viktig foregangsfigur også for den moderne dokumentasjonslingvistikken. I tillegg retter han oppmerksomhet mot dokumentasjonsarbeidet av aleutisk i Alaska som professor Knut Bergsland foretok allerede på 1950-tallet (ibid. 166).

Det er grunn til å skille mellom et relativt nytt fagfelt fra 1990-tallet og virksomhet som ble drevet vesentlig mye tidligere for å dokumentere språk. Jeg referer til fagfeltet med termen «dokumentasjonslingvistikk». Fagfeltet ble etablert på 1990-tallet, og det setter søkelys på dokumentasjon av truede språk. Woodbury (2011: 185 fotnote1) påpeker at også dokumentasjonslingvistikk kunne referere til lingvistisk arbeid der man dokumenterer språk i mer generell betydning. Derfor kunne for eksempel korpuslingvistikk høre til dokumentasjonslingvistikk. Siden dette fagfeltet i dag referer til dokumentasjon av truede språk, bruker han likevel begrepet dokumentasjonslingvistikk i denne etablerte betydningen. Jeg bruker begrepene *dokumentasjonslingvistikk* og *språkdokumentasjon* på samme måte som Woodbury (2011), og referer med språkdokumentasjon til virksomhet som jeg oppfatter å være mye eldre enn fagfeltet som man i dag kaller dokumentasjonslingvistikk. I denne artikkelen vil jeg presentere spesielt hvordan språkdokumentasjon har foregått i finsk språkforskning eller fennistikk.

På 1800-tallet kjente man ikke til finske dialekter, og derfor ble det ansett som viktig å dokumentere dem (Häkkinen 2008: 101–102; Luodonpää-Manni og Ojutkangas 2018: 416–417). Denne virksomheten hører til det man kaller tradisjonell dialektologi. En samfunnsmessig målsetning for innsamling av finske dialekter på 1800-tallet var å få etablert et moderne skriftspråk. En teoretisk målsetning derimot var språkhistorisk på grunn av at det manglet gammel skriftlig dokumentasjon for å kunne belyse språkhistorisk utvikling for finsk (Hovdhaugen et al. 2000: 422). Den sistnevnte målsettingen hadde i tillegg sammenheng med den tidsaktuelle junggrammatiske språkteorien på slutten av 1800-tallet. Målsettingen preget hvordan innsamlingsarbeidet ble gjennomført: man var interessert bare i de eldste talerne av lokale dialekter. Man oppfattet lokale dialekter som homogene, og festet til å begynne ikke oppmerksomhet ved variasjon mellom ulike talere (Hovdhaugen et al. 2000: 422; Palander 2000: 436; Nuolijärvi og Sorjonen 2005: 11–12).

I denne artikkelen refererer språkdokumentasjon til innsamling av alt tilgjengelige språkmateriale. Språkdokumentasjon kan innbefatte trykte tekstsamlinger, ordbøker og grammatiske verk. Disse tre – grammatikk, tekstsamling og ordbok – utgjør en grunnstamme for språkdokumentasjon, slik allerede Boas hadde påpekt (Woodbury 2011: 163; Jalava og Sandman 2018: 601).

[2] "…creation, annotation, preservation, and dissemination of transparent records of a language" Woodbury (2011: 159). Oversettelse fra engelsk av forfatteren.

Noen truede språk er dokumentert i svært liten grad, andre ikke i det hele tatt. Enkelte truede språk er imidlertid dokumentert kun gjennom arkivsamlinger av ulikt språkmateriale (Spence 2018: 179). Selv om språkmateriale i arkiv er en viktig ressurs i arbeidet med å revitalisere språk, er dette også en ressurs som kan være utfordrende å nyttiggjøre. Språklige samlinger i arkiv kan ha vært produsert bare for språkforskning. Forskere har i slike tilfeller vært sentrale i å legge premissene for hvilket materiale som har blitt samlet inn, og på hvilke måter dette har blitt gjennomført. Arkivsamlinger som er produsert av lekmenn, kan være mindre pålitelige, for eksempel pga. at de kan være notert ned med et inkonsekvent skrivesystem eller på en måte som ikke representerer riktig uttale eller språkbruk. Arkivmateriale kan være vanskelig å lokalisere, og innhenting av informasjon fra arkivmaterialet kan være en tidskrevende prosess. Transkripsjoner av språkmateriale kan dessuten være vanskelig å forstå uten lingvistisk utdanning (Spence 2018: 179–180; Austin 2020: 203–205).

Spence (2018: 183–184) påpeker også at arkivmateriale ikke inkluderer alle sider av språklige uttrykk. Muntlig språkbruk er ikke nødvendigvis inkludert i arkivmaterialet dersom det ikke finnes lydopptak. Dokumentasjon av fenomener som er typisk for det talte språket, finner man kun dersom det finnes et stort antall lydopptak av høy kvalitet i arkivene. Også betydningen av ord kan være vanskelig å avgjøre dersom ordene er hentet fra et lite korpus. Som eksempel nevner Spence at tekster utgitt i 1923 på wailaki (et urfolkspråk i California), inneholder fire forskjellige ord som tilsvarer det engelske ordet *to eat* 'å spise'. Er alle disse ordene synonymer, eller har de forskjellige betydninger? Dette er det ikke enkelt å finne ut av når størrelsen på korpuset er bare noen få tusen ord.

Mangler i arkivert språkmateriale kan eventuelt bøtes på for eksempel gjennom å sammenligne språklige trekk med bedre dokumenterte slektsspråk. Om man har en grammatikk tilgjengelig, kan man også rekonstruere språkformer. Det er likevel alltid viktig å tilpasse materiale funnet i arkivene slik at de kan benyttes i revitaliseringen (Austin 2020: 205). For eksempel må lingvistiske transkripsjoner som regel forenkles, og arkivmateriale må organiseres slik at også andre enn lingvister kan ha tilgang til dem (Spence 2018: 183–185).

3. Kvenske dialekter i fennistiske innsamlingsprosjekter

I dette kapittelet gir jeg først en oversikt over forskningshistorie av dialekter i Finland, og etter dette presenterer jeg arkiver der kvensk språkmateriale bevares.

3.1. En kort innføring i forskningen av finske dialekt

Den systematiske innsamlingen av og forskningen på finske dialekter startet på 1870-tallet. Denne forskningsinteressen var til å begynne med en prosess som hadde som mål å utvikle finsk som et skriftspråk som kunne brukes på alle samfunnsområder i Finland (Hovdhaugen et al. 2000: 422; Mielikäinen 2017: 355–356). Forskningen på finske dialekter hørte dermed til den finske nasjonsbyggingen. Etableringen av finsk som skriftspråk spilte en vesentlig rolle i denne prosessen.

Tanken om å utvikle finsk til et moderne nasjonalt skriftspråk hentet sin ideologi fra nasjonalromantikken. Finsk ble et skriftspråk under reformasjonen på 1540-tallet, da det Nye testamente og andre religiøse skrifter ble utgitt på finsk. Under den svenske tiden ble finsk likevel lite brukt i andre skriftlige sjangre enn den religiøse (Häkkinen 1994; Häkkinen 2015). Etter 1809 var den politiske situasjonen endret. Finland var nå et autonomt storhertugdømme som hørte til Russland. Den nye finske nasjonen søkte en historisk identitet gjennom folkediktning og historisk språkforskning som beviste at finsk var i slekt med en rekke språk som ble snakket i Russland (Tommila 1989: 53). På begynnelsen av 1800-tallet økte interessen for innsamlingen av finske folkedikt, noe som ledet til arbeidet med å gi ut folkeeposet Kalevala i 1935. Det finske språket fikk etter hvert en viktig symbolverdi som nasjonens språk. Til å begynne med ønsket man å benytte svensk som administrasjonsspråk selv om Finland ikke lenger var del av Sverige, og svensk var blitt et minoritetsspråk i landet (Häkkinen 1994).

Under svensketiden var universitetet (grunnlagt i 1640) i Finland i Åbo, men hovedstaden i det nye autonome Finland ble plassert i Helsingfors. Etter at Åbo brant ned i 1827, ble universitetet flyttet til Helsingfors. En viktig begivenhet for utviklingen av faget finsk ved universitetet var etableringen av et lektorat i finsk i 1829. Det første professoratet i finsk ble etablert i 1851. Den første professoren i finsk

språk og litteratur i Helsingfors ved Kejserliga Alexandersuniversitetet i Finland het Matias Aleksander Castrén som er bedre kjent som forsker av finsk-ugriske språk – spesielt de samojediske språkene – og er kjent for å ha grunnlagt sammenlignende uralistikk. Professoratet dekket ikke bare finsk språk og litteratur, men også beslektede språk med tilhørende litteratur og etnografi. Castrén døde allerede i 1852, og etter dette ble Elias Lönnrot valgt til denne stillingen i 1854 (Häkkinen 2008: 96). Lönnrot ble fulgt av August Ahlqvist i 1863. Under Ahlqvist sin tid hørte forskningen av andre finsk-ugriske språk fortsatt til det samme fagområdet. Fennistikk ble etablert som et eget fagfelt først etter at Alhqvist gikk av. Det nye faget ble delt mellom Arvid Genetz, som ble professor i fennougristikk i 1891, og Emil Nestor Setälä, som ble professor i finsk 1893 (Häkkinen 2008: 127).

Mye av forskningen på det finske språket ble foretatt også utenfor universitet. En viktig aktør var «Suomalaisen Kirjallisuuden Seura» *Finska litteratursällskapet* (FLS[3]) som ble grunnlagt i 1831 (Häkkinen 2008: 74–78). Denne foreningen hadde som formål å utvikle finsk til et skriftspråk som kunne brukes på alle samfunnsområder, samt å støtte utgivelsen av finskspråklig litteratur (Sulkunen 2004: 24–27; Häkkinen 2008: 79; Kolehmainen 2014: 58–59). Ved FLS ble det foreslått å samle inn finske dialekter allerede i 1847.

På 1850-tallet begynte FLS å orientere seg mer mot vest enn øst. Enkelte fennomaner – forsvarere av det finske språket – tok avstand fra Russland og orienterte seg mer mot Sverige og allmenneuropeiske nasjonale bevegelser (Sulkunen 2004: 115–116). Spesielt Yrjö Koskinen, en kjent ungfennoman, ønsket å endre fokuset FLS hadde hatt på folkediktning og slektspråk i Russland, og foreslo forskning på språk og kultur hos finner som bodde i Sverige og Norge (Sulkunen 2004: 145). Allerede på 1820-tallet hadde Carl Axel Gottlund – som ble lektor i finsk ved universitetet i Helsingfors i 1839 – foretatt studiereiser i Finnskogen i Norge og Värmland i Sverige (Haugen 2021). Det første prosjektet ved FLS var å sende David Skogman for å samle inn historiske tradisjoner, sanger, eventyr og språkidiomer hos kvener og tornedalsfinner i 1865 (Sulkunen 2004: 147–148).

Professor i finsk, August Ahlqvist fikk etter hvert et anstrengt forhold til ungfennomaner og spesielt til Yrjö Koskinen. Fennomaner ved FLS var blitt delt i ulike fraksjoner på 1860–70-tallet, da en gruppe som ble kalt kulturfennomaner skilte seg ut fra ungfennomaner. August Ahlqvist hørte til den førstnevnte fraksjonen, og Yrjö Koskinen til den andre. Fraksjonene skilte seg i synet på hvordan man skulle betrakte den svenske arven og det svenske språkets status i Finland. Kulturfennomaner hadde en mer positiv holdning til svensk og mente at Finland burde være takknemlig for den svenske arven som hadde bragt vestlige institusjoner og tradisjoner til landet. Ungfennomaner, som var språknasjonalister, mente at svensk i Finland hindret utviklingen av det finske språket (Paunonen 1976: 311; Virtanen 2002: 86–89; Engman 2016: 84–90, 115–118).

Resultatet ble at Ahlqvist stiftet en ny forening i 1876 «Kotikielen seura» «*Forening for finske studier»* (FFS), en forening for studenter og forskere som også skulle arbeide med forskningen av finsk og etableringen av finsk som skriftspråk, slik som FLS også gjorde (Kolehmainen 2014: 59). Forskningen på dialekter samt navneforskning hørte til foreningens program helt fra begynnelsen. Ahlqvist var den første som organiserte en systematisk innsamling av dialekter, som man på dette tidspunktet fortsatt kjente dårlig (Paunonen 1976: 380–381; Häkkinen 2008; Hovdhaugen et al. 2000: 246).

Dialektforskningen skilte seg etter hvert fra det ideologiske arbeidet med å etablere finsk som et skriftspråk, og utviklet seg til å bli et eget forskningsfelt. Setälä etablerte det junggrammatiske forskningsparadigmet i Finland, som ble toneangivende ikke minst innen dialektforskningen (Korhonen 1986: 129; Hovdhaugen et al. 2000: 180, 423–424). Blant de første monografiene om finske dialekter finner vi flere som er skrevet om nordfinske dialekter. Den første er om tornedalsdialekten, og ble skrevet av Paavo Salonius i 1881. Knut Cannelin skrev om kemidialekten i 1888/1889. Begge disse fikk stipend fra FLS for å samle inn dialektmateriale fra disse områdene (Korhonen 1986: 95, 168; Sulkunen 2004: 187).

Martti Airila skrev en avhandling om tornedalsdialekten i 1912. Airila var Setäläs elev, og var på samme måte som Setälä en junggrammatiker. Hans avhandling er en lydhistorisk gjennomgang av denne dialekten, og han sammenligner den med andre finske dialekter. I motsetning til tidligere nevnte forskere,

[3] Jeg bruker det svenske navnet for finske institusjoner når institusjonen har et offisielt svensk navn. Ellers har jeg oversatt navnene til norsk.

sammenligner Airila tornedalsdialekten også med kvenske dialekter. Airila kjente for eksempel til Konrad Nielsens undersøkelse av kvenske dialekter (Airila 1912; Nielsen, manuskript).

Etter hvert ble flere arkiver etablert for å ta vare på innsamlet dialektmateriale. Det eldste arkivet er «Suomen murteiden sana-arkisto» *Finska dialektarkivet*, som har sine røtter fra slutten av 1800-tallet. «Nimiarkisto» *Namnarkivet* ble etablert i 1915. «Suomen kielen nauhoitearkisto» *Finska bandarkivet* ble etablert i 1959 og «Muoto-opin arkisto» «*Arkivet for morfologi*» i 1967, begge ved Helsingfors universitet. Arkivene ble flyttet til «Kotimaisten kielten keskus» *Institutet för de inhemska språken* på 1970-tallet.

3.2. Finska dialektarkivet

Formålet med å samle inn ordforråd i dialekter var å videreutvikle det finske skriftspråket (Strandberg 2004: 13). Allerede på slutten av 1860-tallet hadde man planer om at samlinger av dialektord, setninger og fraseologi skulle lede til utgivelsen av en dialektordbok (Strandberg 2004: 14; Grünthal 2014: 243).

I 1884 foreslo E. N. Setälä som da var formann i FFS, et program for innsamling av ordforrådet i dialekter som også skulle omfatte finskspråklige dialektområder i Nord-Norge og Värmland i Sverige. I 1896 presenterte Setälä sitt ordbokprogram for FLS, og dette året regnes som begynnelsen av Finska dialektarkivet. Dialektordboka skulle inkludere alle ord man kunne finne i allmuespråket, ikke bare såkalte dialektord. I tillegg skulle dialektordboka gi informasjon om bøyningsmønster av ord, og belyse ordbruk i setninger og fraser. Dialektordboka skulle gi et fullstendig bilde av dialekter ikke bare i Finland, men også i Nord-Sverige, Nord-Norge og Värmland (Strandberg 2004: 20).

Ordboksprosjektet ved FLS utviklet seg så sakte at det i 1916 ble stiftet et aksjeselskap for å ivareta samlingen av dialektord (Strandberg 2004: 44). I 1925 ble virksomheten flyttet til en egen statsfinansiert stiftelse (Grünthal 2014: 245). Denne stiftelsen «Sanakirjasäätiö» var en viktig organisasjon for fennistisk forskning gjennom mange år. Flere av dem som arbeidet der, fikk senere universitetsstillinger innen finskfaget. Også studenter og andre interesserte deltok i stiftelsens arbeid (Häkkinen 2008).

I tillegg til lingvister, deltok også stipendiater og andre interesserte i innsamlingen av dialektord allerede på slutten av 1800-tallet (Lappalainen og Siiroinen 1999: 566–567). For at de skulle ha ferdighetene som krevdes for å samle inn ordforråd, ble det utviklet ulike hjelpemidler. En viktig kilde til finsk ordforråd på 1800-tallet var Lönnrots svensk-finsk ordbok som ble gitt ut i to deler i 1874 og 1880 (Paunonen 1976: 369–374; Strandberg 2004: 16). Lönnrots ordbok var den største ordboka på dette tidspunktet når det gjaldt antall finske ord: ordboka inkluderer 200 000 oppslagsord. Ordartiklene er skrevet på svensk. I ordboka ble det tatt med mange østfinske dialektord, men ordboka presenterte også mange avledninger som Lönnrot hadde konstruert selv. Hvilke dialektområder ordene ble samlet inn fra, er ikke nevnt i Lönnrots ordbok (Häkkinen 1994: 420; Strandberg 2004: 7–8).

Lönnrots ordbok ble grunnlaget for innsamlingen av dialektord i finske dialekter. I 1899 ga E.A. Ekman ut «Suomen kielen keräilysanasto» Det finske språkets innsamlingsordliste. Heftet tok, i tillegg til Lönnrots ordbok, også hensyn til avhandlinger som var gitt ut om finske dialekter, for eksempel Cannelins avhandling om kemidialekten, og andre tekster som inneholdt ordforråd av finske dialekter. I forordet til Ekmans innsamlingsordliste ble det fremmet et ønske om at heftet kunne brukes i innsamlingen av finske dialekter ikke bare i Finland, men også utenom Finlands område, som for eksempel i Norge (Strandberg 2004: 26).

I dag omfatter Finska dialektarkivet 8 millioner kartotekkort med 400 000 ord. I årene 1927–90 ble et blad som het «Sanastaja» Ordsamler brukt i arbeidet med frivillige informanter (Häkkinen 2008). Selv om Nord-Norge var inkludert i innsamlingen av dialektord, vet man ikke om det var kvener blant de frivillige innsamlerne.

Enkelte lingvister samlet inn ordforråd i Norge. Blant disse var ekteparet Lyyli og Martti Rapola, som samlet inn kvenske dialekter på 1930-tallet. Martti Rapola var professor i finsk ved Åbo universitet 1924–1930 og ved Helsingfors universitet 1930–57. Han arbeidet ved den tidligere nevnte ordbokstiftelsen der han ledet innsamlingsarbeid av ordforråd i perioden 1920–23. Han var bl.a. interessert i språkhistorie, og gjorde innsamlingsarbeid i Nordreisa og Lyngen i 1935 for å samle inn kvensk ordforråd. Han ga ut to artikler i 1939 og 1940, basert på disse dialektene i FFS sitt tidsskrift *Virittäjä* (Anttila et al 1995: 392).

Lyyli Rapola arbeidet som stipendiat for ordbokstiftelsen og samlet inn dialektord 1927–1933. Hun samlet inn stedsnavn både fra Värmland i Sverige og Finnskogen i Norge. Hun deltok på innsamlingsreise til Lyngen og til Nordreisa med Martti Rapola (Tiedenaisia). Hennes innsamling av kvenske ordforråd inkluderer 1475 ord fra Lyngen og 582 ord fra Nordreisa (Utvik 1996: 41). Senere på 1940-tallet arbeidet Lyyli Rapola ved Namnarkivet i Helsingfors.

Andre som har levert kvensk ordforråd til Finska dialektarkivet, er for eksempel Anna-Riitta Lindgren som arbeidet som stipendiat i Nord-Norge på slutten av 1960-tallet og 1970-tallet.

3.2. Namnarkivet

Også Namnarkivet hører til de eldste språkarkivene i Finland. Allerede i 1876 ble det foreslått av FFS at man skulle begynne å samle inn stedsnavn (Itkonen 1997: 14). Arkivet ble grunnlagt i 1915, da «Tieteellisten seurain paikannimitoimikunta» Stedsnavnkomité for vitenskapelige foreninger ble etablert. Også historiske foreninger, slike som «Suomen muinaismuistoyhdistys» Finska fornminnesforening, var interesserte i navnegransking. I deres program gikk man inn for å samle inn bare spesifikke navn, slik som navn med uklart semantisk innhold eller navn som kunne belyse bosetningshistorie (Uusitalo 2015).

Stedsnavnkomitéen hadde derimot som mål å arrangere systematisk innsamling av finske og samiske stedsnavn i Finland, inkludert nærområder i tilstøtende land. Prinsipper om hvilke stedsnavn som skulle tas med i samlingen, ble utvidet, og det ble bestemt at også navn på mindre steder og navn som hadde en tydelig semantisk betydning, skulle være inkludert (Uusitalo 2015). Dette prinsippet nevner Kiviniemi (1990: 14) som et av de viktige prinsippene for å legge grunnlaget for forskningen på stedsnavn i Finland. Han skriver at den største delen av samlingen kommer fra 1950-tallet (ibid. 27). Kvenske stedsnavn som finnes i Namnarkivet i Finland er samlet inn av finske forskere på 1970- og 80-tallet. Bland disse var Anna-Riitta Lindgren, Riitta Matilainen, Outi Honkasalo og Eira Söderholm (Kvenske stedsnavn).

Namnarkivet inneholder over 2,7 millioner navn som er registrert i Finland eller i Finlands nærområder, blant annet i Nord-Norge (Itkonen 1997: 19). I Namnarkivet finnes ca. 12 400 navnesedler med kvenske stedsnavn som er innsamlet i Sør-Varanger, Vardø, Vadsø, Tana, Nesseby, Alta, Kvænangen, Kåfjord, Nordreisa, Storfjord og Lyngen. Kvensk stedsnavnstjeneste har kopier av disse samlingene (Andreassen 2015: 88).

3.3. Finska bandarkivet

Finska bandarkivet ble grunnlagt i 1959 ved Helsingfors universitet. Initiativtaker var professor Pertti Virtaranta som hadde fått idéen om et arkiv med lydopptak for dialekter mens han arbeidet i Sverige som universitetslektor i finsk (Yli-Paavola 1970: 10). De fleste materialene i arkivet er båndopptak av finske dialekter som ble systematisk samlet inn på 1960- og -70-tallet (Suutari 2010). Innsamlingen resulterte i at Finland trolig har den største samlingen av dialektmaterialer på bånd i hele verden (Hovdhaugen et al. 2000: 425). Samlingene inkluderer også dialekter på kvensk og meänkieli, som man på innsamlingstidspunktet oppfattet som finske dialekter.

Finska bandarkivet søkte om nordiske midler til innsamlingsarbeidet av kvenske dialekter på 1970-tallet uten å motta støtte. Å samle inn kvenske dialekter til arkivets dialektsamlinger ble imidlertid ansett for å være så viktig at arkivet bestemte seg for å selv finansiere innsamlingsreisene til Norge, fordi man opplevde dette som arkivets ansvarsområde (Lyytikäinen 1982: 151). Dette forteller at finske språkforskere fortsatt på 1960- og -70- tallet hadde en oppfatning om at innsamlingen av kvenske dialekter hørte til det fennistiske dokumentasjonsprosjektet.

Allerede i 1959 ble de første opptakene av kvenske dialekter gjort (Kvensk institutt). På 1960-tallet samlet enkelte stipendiater inn kvenske dialekter. Pekka Laaksonen samlet inn 30 timer med dialektopptak i Vadsø-området. Stipendiatene Marjut Aikio og Anna-Riitta Lindgren samlet inn 30 timer kvenske dialekter i Nordreisa (Lyytikäinen 1982: 160). I tillegg til båndopptak, samlet de også andre typer materiale med kvenske dialekter. Lindgren (2014) beskriver sin tid som stipendiat og sitt arbeid for å samle inn kvenske dialekter i Nordreisa i sin artikkel i Ottar.

Lyytikäinen (1982) gir en oversikt over hvordan Finska bandarkivet systematisk organiserte innsamlingsreiser til områder der man fortsatt kunne finne kvener som snakket sitt språk på 1970-tallet. Det

ble gjort flere reiser til Nord-Norge i 1971–1973. Intervjuerne var Erkki Lyytikäinen, Juhani Pallonen, Matti Punttila, Jorma Rekunen, Pentti Soutkari og Jaakko Yli-Paavola.

Opptakene av kvenske dialekter ble samlet i et stort geografisk område fra Varanger i Øst-Finnmark til Storfjord i Troms, og alle bygdene der det bodde kvener, inkludert mange små steder, ble besøkt. Lyytikäinen (1982: 159) konstaterer at disse opptakene gir et godt bilde av hvordan kvenske dialekter ble snakket på 1970-tallet. Finska bandarkivet hadde sammenlagt 374 timer opptak av kvenske dialekter i slutten av 1981. De mest omfattende opptakene kommer fra Vadsø (80 timer), Nordreisa (73 timer), Sør-Varanger (50 timer), Alta (40 timer) og Porsanger (37 timer) (Lyytikäinen 1982: 160). Senere er det gjort flere opptak, slik at arkivets kvenske dialektopptak teller totalt over 400 timer.

Transkripsjoner av kvenske dialekter bruker finsk-ugrisk fonetisk alfabet (SUT). Denne transkripsjonen ble utviklet av Setälä ved hjelp av den svenske språkforskeren J.K.Wiklund, og den ble presentert i 1901 (Häkkinen 2008). Spesielt for dette fonetiske alfabetet er at man bruker enkelte spesifikke fonetiske tegn som er annerledes enn i det internasjonale fonetiske alfabetet (IPA). Selv om man har forenklet SUT gjennom årene til det man kaller en halvgrov transkripsjon (Yli-Paavola 1970: 75–79; Iivonen et al. 1990: 53–60), kan det likevel være utfordrende å lese transkripsjoner av kvenske dialekter dersom man ikke er kjent med slik transkripsjon. Se for eksempel Spence (2018: 179) om viktigheten å forstå hvordan arkiverte materialer er blitt transkribert.

3.4. «Arkivet for morfologi»

Finske arkiver inneholder også samlinger av morfologi, dvs. informasjon om språkets bøyningsformer og orddanning. «Arkivet for morfologi» ble grunnlagt ved Universitetet i Helsingfors i 1967. Som mål hadde forskerne å samle inn morfologi i hvert tredje finskspråklige sogn, slik at hele språkområdet ble dekket (Itkonen 1997: 13).

Samlingene av morfologi fulgte en innsamlingsplan der målet var å dekke bøyningskategorier og orddanningsprinsipper i finsk. Hver samling inneholder ca. 2500–3000 arkivkort, og det tok 4–6 måneder for de stipendiatene som deltok i arbeidet, å ferdigstille en samling. Innsamlingen av morfologi ble primært gjort ved å bruke opptak av spontan tale, men også gjennom å spørre informanter om ulike former (Juusela 1987: 297, 302). Arkivet ble digitalisert på 1990-tallet.

Arkivet inneholder samlinger av kvenske dialekter fra Nordreisa og Øst-Finnmark. Disse er samlet inn av Anna-Riitta Lindgren og Marjut Aikio på 1960- og -70-tallet. Samlingen fra Nordreisa er den mest omfattende. Begge disse samlingene er tilgengelige i digital form.

Samlingene i Nordreisa ble samlet inn i 1969–70 da forskerne oppholdt seg i Nordreisa i lange perioder. Kvensk var fortsatt i bruk som dagligtale i Nordreisa i ca. 10 husholdninger på dette tidspunktet. Informantene var født i perioden 1882–1926 (Lindgren 1972).

Arkivmaterialet om kvensk er også brukt i forskningen på kvensk. Lindgrens avhandling (1993) var den første doktoravhandlingen om kvensk, og den baserer seg bl.a. på den samlingen med morfologi som hun selv var med på å produsere. Også andre forskere har brukt kvensk materiale som finnes i finske arkiver.

4. Arkivenes betydning for kvensk revitalisering

Som denne oversikten viser, finnes det mye språklig dokumentasjon av kvensk i ulike finske arkiver. Når det i tillegg finnes dokumentasjon av kvensk i Norge, kan man konstatere at kvensk er et minoritetsspråk som er rimelig godt dokumentert. I arkivene finnes også ulike typer materialer som øker mulighetene for å bruke arkivmaterialet i revitaliseringsarbeidet.

4.1. Etablering av kvensk skriftspråk

Et sentralt mål i revitaliseringsarbeidet av kvensk er å utvikle kvensk skriftspråk. En viktig motivasjon bak dette var behovet for å produsere læremidler i kvensk. Dette arbeidet begynte etter at kvensk ble anerkjent som språk i 2005, og Kvensk institutt fikk ansvar for arbeidet. Den skriftlige standarden er basert på muntlig kvensk. For å kunne gjennomføre dette arbeidet, overtok Kvensk institutt 419 timer opptak av kvenske dialekter fra Finska bandarkivet. Halvparten av materialet var allerede digitalisert da Kvensk institutt mottok det, og resten har vært digitalisert ved instituttet. Opptakene oppbevares i dag ved Kvensk institutt.

Materialet er brukt i instituttets eget språkarbeid, men også slektninger til de som ble intervjuet, kan få tilgang til opptakene (Kvensk institutt). På sidene til Institutet för de inhemska språken i Helsingfors kan man få informasjon om hvilke av de kvenske dialektopptakene som er transkriberte, og interesserte forskere kan søke om tillatelse til å studere disse dialektopptakene (Kotus).

En viktig begivenhet i arbeidet med å etablere kvensk som skriftspråk, var utgivelsen av kvensk grammatikk av Eira Söderholm. Grammatikken kom ut på kvensk i 2014, og på norsk i 2017. Forfatteren nevner opptakene som ble gjort av Finska bandarkivet, som den viktigste kilden for arbeidet med grammatikken. Også samlinger av kvensk morfologi ved Arkivet for morfologi ble brukt som kilde til grammatikken (Söderholm 2017: 28–29). Kvensk grammatikk følger avgjørelser som Kvensk språkting – som består av kvenske språkbrukere – har foretatt når det gjelder valg mellom ulike mulige alternativer i standardiseringen av kvensk (Keränen 2018: 8–10).

Söderholms grammatikk er en referansegrammatikk (se Jalava og Sandman 2018: 600) i kvensk som gir en oversikt over kvensk fonologi, morfologi og syntaks. Kvensk grammatikk presenterer den såkalte læreboknormalen og har som målsetting å hjelpe språkbrukere som ønsker å skrive kvensk og dem som ønsker å lære seg kvensk (Söderholm 2017: 37; Keränen 2018: 10).

Kvensk grammatikk har vært avgjørende også i arbeidet med utviklingen av kvenske digitale verktøy på internett (Giellatekno). Spesielt viktig var Kvensk grammatikk for arbeidet å lage en morfologisk analysator, som er knyttet til en digital kvensk ordbok. Den kan hjelpe språkbrukere også med bøyninger av kvenske ord (Haavisto et. al. 2014: 181). I dette arbeidet har spesielt Trond Trosterud gitt et verdifullt bidrag.

Finsk dialektordbok baserer seg på Finska dialektarkivet. Dialektordboka er ikke ferdigstilt i sin helhet ennå, til tross for at det er mer enn hundre år siden Setälä presenterte de første planene for denne ordboka. Dialektordboka kom ut i bokform i 8 deler etter 1985, men i 2012 ble det bestemt at hele dialektordboka skal gis ut i digital form (SMS). For etablering av kvensk skriftspråk er det behov for å hente ord fra muntlig kvensk (Söderholm 2005), og slik ord er også dokumentert i den finske dialektordboka, i tillegg til at kvenske ord også kan hentes fra Finska dialektarkivet og dialektopptakene i Finska bandarkivet. Derfor kan det fennistiske dokumentasjonsarbeidet med å samle inn ordforråd som også inkluderte kvenske dialekter, få betydning for utviklingen av ordforrådet i kvensk skriftspråk.

Man antar ofte at utvikling av en skriftstandard er vesentlig for revitaliseringen. Uten skriftspråk er det vanskelig for eksempel å lage læremateriell for undervisning. Likevel kan en skriftlig varietet av et minoritetsspråk bli møtt med skepsis av språkbrukere som er vant til å bruke språket bare muntlig. Skriftspråket kan virke fjernt for eksempel fordi det inneholder former som man ikke kjenner fra egen dialekt. Språkbrukere kan derfor også avvise en skriftlig standard basert på sitt minoritetsspråk (Lane 2015: 280). På grunn av dette inkluderer man en god del variasjon i kvensk skriftspråk, slik at flere språkbrukere kan kjenne igjen muntlige former som de bruker (Söderholm 2017). Selv om bruk av skriftlig kvensk har økt de siste årene, er det fortsatt store utfordringer knyttet til å spre bruken av den på flere områder. Etableringen av kvensk som skriftspråk har likevel allerede gjort at det gis mer oppmerksomhet til kvener som minoritet enn tidligere (Keränen 2018: 12–14).

4.2. Bruk av kvenske stedsnavn

Kvenske stedsnavn presenteres i digital form på databasen «Kvensk stedsnavndatabase», som har 10 000 registrerte stedsnavn. En god del av originalmaterialet kommer fra Namnarkivet i Institutet för de inhemska språken i Helsingfors. I tillegg er flere kvenske stedsnavn registrert i Sentralt stedsnavnregisteret i Norge (SSR). I Kvensk stedsnavndatabase kan man søke på stedsnavn, og det er mulig å finne forklaring av det kvenske navnet på norsk, i tillegg til parallellnavn på samisk og norsk (Kvenske stedsnavn).

Den norske stedsnavnloven fra 1990 forutsetter at kvenske stedsnavn skal tas vare på, og myndighetene skal medvirke til kjennskap til navnene og aktiv bruk av dem. Siden 2015 har kvensk ortografi blitt brukt i stedsnavnene (Målrettet plan 2017–2021, 21). Flere kvenske stedsnavn er skiltet enn tidligere. I 2019 ble det satt opp 50 nye skilt som inneholdt stedsnavn på kvensk, i tillegg til norsk og samisk (Lanes 2019). Bruk av minoritetsspråklige stedsnavn på offentlige skilt bidrar til synliggjøring av minoritetsspråk i det offentlige rom.

4.3. Fennistisk språkdokumentasjon sammenlignet med dokumentasjon i dokumentasjonslingvistikk

At kvensk tidlig ble inkludert i fennistiske innsamlingsprogrammer, har altså hatt betydning for revitaliseringen av kvensk i dag. Dette kan virke som et paradoks siden fennomanenes iver for å inkludere områder utenfor Finlands grenser i sine prosjekter, vekket bekymring i Norge for at Finland hadde politiske interesser overfor de områdene hvor det var bosatt kvener (Ryymin 2004). Så lenge kvensk ble ansett for å være en finsk dialekt, var det naturlig å inkludere kvenske dialekter i fennistiske prosjekter. Språkforskere som deltok i prosjektene, hadde også vitenskapelig interesse av å hente frem all tilgjengelig informasjon om det finske språket. Dokumentasjon av dialekter ble ansett som viktig også på grunn av at man var redd for at dialekter skulle forsvinne (Strandberg 2004: 51, 55, 74; Nuolijärvi og Sorjonen 2005: 12).

Et av kjennetegnene på det fennistiske vitenskapsparadigmet har vært et sterkt søkelys på empirisk materiale. Karlsson (1975) kritiserte dette, og påpekte at også materialinnsamling er styrt av et vitenskapelig paradigme. For eksempel har diskusjonstemaer i opptakene av dialekter ofte satt søkelys på temaer som hører til gamle levemåter, og valg av tematikk påvirker derfor hvordan materiale kan brukes. Fennistiske materialinnsamlingsprosjekter ble ikke gjennomført med tanke på språkrevitalisering i fremtiden. Derfor har dette materialet selvsagt også mange begrensninger når man ønsker å bruke det i revitalisering av et minoritetsspråk.

Arkivmateriale som er samlet inn ved fennistiske språkdokumentasjonsprosjekter, skiller seg fra dokumentasjon av språk som man gjør i dag i dokumentasjonslingvistikken. Der er målet å samle inn materiale i språkets naturlige omgivelser i ulike kontekster. Diskusjonstemaer skal også inkludere temaer fra hverdagslivet, og ikke bare presentere gamle levemåter. Dokumentasjonen skal inneholde interaksjon mellom språkbrukere, og språkbruk fra alle aldersgrupper er interessant. Språkbrukere skal være engasjert i innsamlingsarbeidet, og samarbeide med forskerne (Austin og Sallabank 2018: 209–211). I tillegg bidrar den teknologiske utviklingen til at det finnes flere muligheter for å dokumentere språk, og også til å presentere og bruke innsamlet språklig materiale, enn da man drev feltarbeid innen tradisjonell dialektologi.

Slik Austin og Sallabank (2011: 212) påpeker, har dokumentasjonslingvistikken hentet ideer fra sosiolingvistikken når det gjelder hvordan dokumentasjon helst skal gjennomføres. Innsamlingen av kvensk materiale ble gjort før sosiolingvistikken ble etablert i fennistikken på 1970-tallet (Nuolijärvi og Sorjonen 2005: 12), noe som betyr at paradigmet som styrte den fennistiske språkdokumentasjonen før denne tiden, som regel var det samme som innenfor tradisjonell dialektologi. Tross dette hadde finske forskeres arbeid med å samle inn kvenske dialekter betydning for minoriteten allerede på 1960–70-tallet. På dette tidspunktet var kvener fortsatt en usynlig minoritet i Norge. Einar Niemi (2010: 47) påpeker at interessen som finske forskere viste for kvener, var en del av prosessen som på 1960-tallet gjorde at man oppdaget den kvenske minoriteten etter en periode hvor man hadde trodd at den var fullstendig assimilert og nærmest ikke eksisterte lenger.

5. Avslutning

Kvenske dialekter har vært inkludert i fennistiske innsamlingsprosjekter helt siden FLS sendte David Skogman til Norge på 1860-tallet. Materialinnsamlingen av kvensk hørte til prosjektet som i første fase var knyttet til fennomaners nasjonsbyggingsprosjekt der utviklingen av et moderne skriftspråk for finsk spilte en viktig rolle. Dialekter ble et eget forskningsobjekt innen fennistikken mot slutten av 1800-tallet. Systematiske innsamlingsprosjekter av dialektmateriale preget det fennistiske vitenskapsparadigmet i en lang periode (Karlsson 1975).

Flere innsamlingsprosjekter har resultert i at finske arkiver disponerer mye materiale av kvenske dialekter. Selv om innsamlingen av dialekter ikke hadde som formål at materialet skulle brukes i språkrevitaliseringsarbeid, har dette materialet likevel vært til hjelp for revitaliseringen av kvensk i dag. Spesielt viktig har intervjumateriale av kvenske dialekter fra 1960–70-tallet vært i prosessen med å utvikle et kvensk skriftspråk.

Spence (2018) påpeker at arkivmateriale ikke kan erstatte levende språkbrukere. Selv om kvensk er et truet språk i dag, finnes det fortsatt mange personer som kan snakke språket. Språkrevitalisering er krevende arbeid. En del viktige forutsetninger er imidlertid på plass med revitalisering av kvensk.

Litteratur

Airila, Martti. 1912. *Äännehistoriallinen tutkimus Tornion murteesta murteen suhdetta muihin murteihin silmällä pitäen*. Suomi IV:12, 1–244. Suomalaisen Kirjallisuuden Seura, Helsinki.

Andreasssen, Irene. 2015. Kvenske stedsnavn i Norge. I *Namn i det fleirspråklege Noreg*, red. av Gulbrand Alhaug og Aud-Kirsti Pedersen, s. 85–103. Novus forlag, Oslo.

Anttila, Raimo, Kaisa Juusela og Heikki Paunonen. 1995. Miten muodot muuttuvat? Ensimmäinen väitös suomen kielestä Norjassa. *Virittäjä* 3: 390–408.

Austin, Peter K. 2020. Language Documentation and Language Revitalization. I *Revitalizing endangered languages: a practical guide*, red. av Justyna Olko og Julia Sallabank, s. 199–212. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108641142.014

Austin, Peter og Julia Sallabank. 2018. Language Documentation and Language Revitalization. Some Methodological Considerations. I *The Routledge Handbook of Language Revitalization*, red. av Leanne Hinton, Leena Huss og Gerald Roche, s. 207–215. Routledge, New York. https://doi.org/10.4324/9781315561271-26.

Engman, Max 2016. *Språkfrågan. Finlandsvenskhetens uppkomst 1812–1822*. Svenska litteratursällskapet i Finland, Helsinki.

Giellatekno. Språkverktøy for kvensk. https://giellatekno.uit.no/cgi/index.fkv.fin.html

Grünthal, Riho. 2014. Setälän suuren sanakirjaohjelman tausta. *Virittäjä* 2: 242–249.

Haavisto, Mervi, Kaisa Maliniemi, Leena Niiranen, Pirjo Paavalniemi, Tove Reibo og Trond Trosterud. Kvensk ordbok på nett - hvem har nytte av den? I *Nordiske studier i leksikografi 12. Rapport fra Konferanse om leksikografi i Norden Oslo 13.–16. august 2013*, red. av Ruth Vatvedt Fjeld og Marit Hovdenak. Skrifter utgitt av Nordisk forening for leksikografi (13), s. 176–192.

Haugen, Morten Olsen. 2021. *Carl Axel Gottlund* i *Store norske leksikon* på snl.no. Hentet 16. april 2022 fra https://snl.no/Carl_Axel_Gottlund.

Hovdhaugen, Even, Fred Karlsson, Carol Henriksen og Bengt Sigurd. 2000. *The History of Linguistics in the Nordic Countries*. Societas Scientiarum Fennica. Gummerus Kirjapaino Oy, Jyväskylä.

Häkkinen, Kaisa. 1994. *Agricolasta nykykieleen. Suomen kirjakielen historia*. WSOY, Helsinki. https://doi.org/10.21435/sflin.19

Häkkinen, Kaisa. 2008. *Suomen kielen historia 2. Suomen kielen tutkimuksen historia*. Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja, Turku.

Häkkinen, Kaisa. 2015. *Spreading the Written Word: Mikael Agricola and the Birth of Literary Finnish*. Suomalaisen Kirjallisuuden Seura, Helsinki.

Hyltenstam, Kenneth og Tommaso Milani. 2003. *Kvenskans status: Rapport for Kommunal- og regionaldepartementet og Kultur- og kirkedepartementet i Norge*. Oslo. https://www.regjeringen.no/globalassets/upload/kilde/kkd/hdk/2003/0003/ddd/pdfv/193348-kvenrapport_hyltenstam_slutversion_oktober.pdf.

Iivonen, Antti, Antti Sovijärvi og Reijo Aulanko. 1990. *Foneettisen kirjoituksen kehitys ja nykytila. Kansainvälinen foneettinen aakkosto (IPA). Suomalais-ugrilainen tarkekirjoitus (SUT)*. Helsingin yliopiston fonetiikan laitoksen monisteita n:o 16. Helsingin yliopisto, Helsinki.

Itkonen, Terho 1997. *Nimestäjän opas. Apuneuvoja suomalais-ugrilaisten kielten opintoja varten XIII.* Kolmas, uusittu painos. Suomalais-ugrilainen seura, Saarijärvi.

Jalava, Lotta og Erika Sandman. 2018. Kielten dokumentointi ja kieliopin kuvaus. I *Kielentutkimuksen menetelmiä I–V*, red. av Milla Luodonpää-Manni, Markus Hamunen, Reetta Konstenius, Matti Miestamo, Urpo Nikanne og Kaius Sinnemäki, 596–638. Suomalaisen Kirjallisuuden Seuran toimituksia 1457. Suomalaisen Kirjallisuuden Seura, Helsinki. https://doi.org/10.2307/j.ctv1qp9hgb.21.

Juusela, Kaisu. 1987. Muoto-opin arkisto kaksikymmenvuotias. *Virittäjä* 91 (3): 289–313.

Karlsson, Fred 1975. Fennistiikan tieteenparadigmasta ja sen ohjausvaikutuksesta. *Virittäjä* 79 (2): 179–193.

Keränen, Mari. 2018. Language Maintenance through Corpus Planning – the Case of Kven. *Acta Borealia* 35 (2): 176–191. https://doi.org/10.1080/08003831.2018.1536187.

Kiviniemi, Eero. 1990. *Perustietoa paikannimistä*. Suomalaisen Kirjallisuuden Seura, Helsinki.
Kolehmainen, Taru. 2014. *Kielenhuollon juurilla. Suomen kielen ohjailun historiaa*. Suomalaisen Kirjallisuude Seura, Helsinki.
Korhonen, Mikko. 1986. *Finno-Ugrian language studies in Finland 1828–1919*. The History of Learning and Science in Finland 1828–1918 11. Societas Scientiarum Fennica, Helsinki.
Kotus. Kotimaisten kielten keskus. Institutet för de inhemska språken. https://www.kotus.fi.
Kvensk institutt. Lydopptak av kvenske dialekter. https://www.kvenskinstitutt.no/eget-sprakarbeid/ressurser/lydopptak-av-kvenske-dialekter/.
Kvenske stedsnavn. Kvensk stedsnavndatabase. http://www.kvenskestedsnavn.no.
Lane, Pia. 2015. Minority language standardization and the role of users. *Language Policy* 14 (3): 263–283. https://doi.org/10.1007/s10993-014-9342-y.
Lanes, Laila. 2019. Over 50 kvenske navn på skilt. NRK 11.11.2019. http://www.kvenskestedsnavn.no.
Lappalainen, Hanna og Mari Siiroinen. 1999. Kotikielen seuran toimintaa sata vuotta sitten. *Virittäjä* 103 (4): 556–571.
Lindgren, Anna-Riitta. 1972. *Keruukertomus*. Morfologian arkisto. Kokoelma saatu arkistoon 25.10.1972. Forfatteren har kopi.
Lindgren, Anna-Riitta. 1993. *Miten muodot muuttuvat. Ruijan murteiden verbitaivutus Raisin, Pyssyjoen ja Annijoen kveeniyhteisöissä.* With English Summary. Doktoravhandling. Universitetet i Tromsø.
Lindgren, Anna-Riitta. 2014. Litt før den etniske renessansen – feltarbeid blant kvener på slutten av 1960-tallet. *Ottar* 5 = 303: 34–40.
Luodonpää-Manni, Milla og Krista Ojutkangas. 2018. Laadullinen aineistopohjainen kielentutkimus. I *Kielentutkimuksen menetelmiä I–V*, red. av Milla Luodonpää-Manni, Markus Hamunen, Reetta Konstenius, Matti Miestamo, Urpo Nikanne og Kaius Sinnemäki, 412–441. Suomalaisen Kirjallisuuden Seuran toimituksia 1457. Suomalaisen Kirjallisuuden Seura, Helsinki. https://doi.org/10.2307/j.ctv1qp9hgb.15.
Lyytikäinen, Erkki. 1982. Nauhoitusmatkalla Ruijassa. I Ulkosuomalaisia red. av Pekka Laaksonen og Pertti Virtaranta. *Kalevalaseuran vuosikirja* 62: 151–165. Suomalaisen Kirjallisuuden Seura, Jyväskylä.
Mielikäinen, Aili. 2017. Ensimmäisten suomenkielisten murretutkimusten metakieli. *Virittäjä* 121 (3): 355–381. https://doi.org/10.23982/vir.58386.
Målrettet plan 2017–2021. Vidare insats for kvensk språk. Kommunal- og moderniseringsdepartementet.
Nielsen, Konrad. Manuskript. *Äänneopillisia seikkoja : kahdesta Norjan suomalaisesta murteesta (muistiinpanojen perusteella).* Håndskrevet manuskript. UiO, Universitetsbibliotek.
Niemi, Einar. 2010. Kvenene – Nord-Norges finner. En historisk oversikt. I *Nasjonale minoriteter i det flerkulturelle Norge*, red. av Anne C. Bonnevie Lund og Bente Bolme Moen, s. 33–50. Tapir akademisk forlag, Trondheim.
Nuolijärvi, Pirkko og Marja-Liisa Sorjonen. 2005. *Miten kuvata muutosta. Puhutun kielen tutkimuksen lähtökohtia murteenseuruuhankkeen pohjalta.* Kotimaisten kielten tutkimuskeskuksen verkkojulkaisuja 13. Näköispainos. Kotimaisten kielten tutkimuskeskus, Helsinki. Tilgjengelig på https://kaino.kotus.fi/www/verkkojulkaisut/julk13/miten_kuvata_muutosta_verkkojulkaisu_13.pdf
Palander, Marjatta. 2000. Puhetutkimuksen uudet haasteet. Virkaanastujaisesitelmä. Joensuun yliopistossa 12. toukokuuta 2000. *Virittäjä* 104 (3): 436–441.
Paunonen, Heikki. 1976. Kotikielen Seura 1876–1976. *Virittäjä* 80 (3–4): 310–432.
Ryymin, Teemu. 2004. *«De nordligste finner.» Fremstillingen av kvenene i den finske litterære offentligheten 1800–1939*. Speculum Boreale 6. Universitetet i Tromsø, Tromsø.
SMS. Suomen murteiden sanakirja. https://www.kotus.fi/sanakirjat/suomen_murteiden_sanakirja.
Spence, Justin. 2018. Learning Languages Through Archives. *The Routledge Handbook of Language Revitalization*, red. av Leanne Hinton, Leena Huss og Gerald Roche, 179–187. Routledge. New York. https://doi.org/10.4324/9781315561271-23
Strandberg, Jan. 2004. *Ei sanat salahan joua: fennistiikan murteenkeruun historiaa 1868–1925*. Pro-gradu tutkielma. Helsingin yliopisto, Suomen kielen laitos. Tilgjengelig på http://urn.fi/URN:NBN:fi-fe20041606.

Sulkunen, Irma. 2004. *Suomalaisen Kirjallisuuden Seura 1831–1892*. Suomalaisen Kirjallisuuden Seura, Helsinki.
Suutari, Toini. 2010. Suomen kielen nauhoitearkisto – vireä viisikymppinen. *Virittäjä*, 114 (3): 423–426.
Söderholm, Eira. 2005. Planlegging av kvensk språk – utvikling av ordforråd. I *Kvener og skogfinner i fortid og nåtid.* Rapport fra seminaret «Kvener og skogfinner i fortid og nåtid identitetsforvaltning og strategier», Vadsø oktober 2005, red. av Anna-Riitta Lindgren, Einar Niemi, Marit Anne Hauan, Leena Niiranen og Trond Thuen. Speculum Boreale 9, s. 43–50. Skriftserie frå Institutt for historie, Tromsø.
Söderholm, Eira. 2017. *Kvensk Grammatikk*. Cappelen Damm akademisk, Oslo. https://doi.org/10.23865/noasp.24.
Tiedenaisia. Vitenskapskvinner. Women of Learning. Lyyli Rapola, nimistöntutkija, paikannimiarkiston hoitaja 1904–1979. Tilgjengelig på https://www.mv.helsinki.fi/home/eisaksso/tiedenaiset/rapola.html
Tommila, Päiviö. 1989. Mitä oli olla suomalainen 1800-luvun alkupuolella. I *Herää Suomi. Suomalaisuusliikkeen historia*, red. av Päiviö Tommila og Maritta Pohls, s. 51–65. Kustannuskiila Oy Kuopio, Jyväskylä.
Utvik, Hanne. 1996. *Norske ord i finsk språkdrakt: en studie av nyere skandinaviske substantivlån i kvensk/ruijafinsk tekstmateriale med hovedvekt på norske lån*. Hovedfagsoppgave. Universitetet i Tromsø.
Uusitalo, Helinä. 2015. 100-vuotias nimiarkisto. Kotuksen kolumnit 9.4.2015. Tilgjengelig på https://www.kotus.fi/nyt/kolumnit_artikkelit_ja_esitelmat/muita_kotuslaisten_kolumneja/100-vuotias_nimiarkisto.17581.news
Virtanen, Matti. 2002. *Fennomanian perilliset: Poliittiset traditiot ja sukupolvien dynamiikka*. Suomalaisen Kirjallisuuden Seura, Helsinki.
Woodbury, Anthony C. 2011. Language Documentation. I *The Cambridge Handbook of Endangered Languages*, red. av Peter K. Austin og Julia Sallabank, s. 159–186. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511975981.009.
Yli-Paavola, Jaakko. 1970. *Vuosikymmen kielennauhoitusta. Suomen kielen nauhoitearkiston toimintaa v. 1959–1968*. Tietolipas 60. Suomalaisen Kirjallisuuden Seura, Vammala.

Low hanging fruit and the Boasian trilogy in digital lexicography of morphologically rich languages: Lessons from a survey of Indigenous language resources in Canada

Elizabeth Pankratz, Antti Arppe, and Jordan Lachler
University of Alberta

Abstract

Online lexicographical resources for the morphologically rich Indigenous languages in Canada use a wide range of strategies for conveying their language's morphological system, i.e. how words are inflected and derived, which this paper illustrates in a survey of seventeen bilingual online resources. The strategies these resources employ boil down to two basic approaches to the underlying structure of the resource: 1) a lexical database, or 2) a computational model. Most resources we surveyed are constructed around lexical databases. These assume the word(form) as the basic unit, an assumption that makes it difficult to incorporate the language's sub-word, morphological structure in full detail. However, one resource uses a computational morphological model to bring the language's morphology into the core of the lexicon – this proved to be a "low-hanging fruit" in the application of language technology that had been accomplished within a reasonable time-frame, as has been advocated by Trond Trosterud. We discuss the value created and questions raised by this approach and argue that it successfully overcomes the traditional Boasian three-way partition of dictionary, grammar, and text, creating integrated language resources that meet the modern needs of low-resource endangered languages and their communities.

Keywords: Indigenous languages, electronic dictionaries, lexicography, morphology, finite-state modeling, Plains Cree, Canada

1 Introduction

Modern learners of majority languages have many high-quality digital resources at their fingertips, like online dictionaries, linguistically analyzed collections of written and spoken language, spell-checkers, grammar-checkers, and language-learning applications. It is easy to take access to these for granted. However, learners of smaller, lower-resourced, and often endangered languages face a different situation. Such languages are much less likely to have high-quality digital resources, even though these are the languages whose need for them is greatest. Strong digital resources make the language more accessible to current speakers and learners, facilitating language revitalization, and they are also valuable for documentation and preserving the language for the future (cf. Trosterud 2006; Arppe et al. 2016: 6).

So that any of these goals can be achieved, resources must accurately reflect the language they encompass. A challenge to this is that the Indigenous languages in Canada are morphologically very rich – some lemmas may have hundreds of different inflections – and yet the resources that are used to convey them are not always structured in such a way that makes incorporating complex morphological systems straightforward.

The *de facto* standard in bilingual lexicography is to compile lists in which words or phrases in the Indigenous language and the majority language are mapped to each other one-to-one (cf. Prinsloo 2012, Lew 2012, Granger 2012). Once digitized, these word or phrase lists are stored as lexical databases, which assume the entire word as the basic unit, informed as they are by the Indo-European majority language tradition, in which this assumption is unproblematic.

However, taking the phonological or orthographic word as a basic unit when working with languages with very complex morphological structure – such as the Indigenous languages of North and South America

 Nordlyd 46.1: 193–204, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6441

– means that lexicographers need to employ more creative strategies to still convey the language's morphology in a clear way. Complex inflectional morphology is usually captured using inflectional paradigms: tables that map certain combinations of morphosyntactic properties (that appear in the rows and columns) onto certain inflected forms (that appear in the cells; cf. Spencer 2016: 27; Stump 2016: 7–10). A full paradigm contains the entire set of inflected forms that are associated with (and are possible for) a particular lemma, i.e. the citation form, or base form, which has been selected to represent the entire lexeme, which is manifested by all these possible inflected forms. However, without great technical effort it is difficult to incorporate inflectional paradigms into a word-based resource – what counts as the head word? How are the various inflected forms linked together? How are multimorphemic inflected forms assembled in the first place? The ways in which online resources for morphologically complex Indigenous languages in Canada approach these and further questions will be explored in what follows.

The current contribution shares the results of a metalexicographical survey of existing online bilingual resources for Indigenous languages in Canada with a focus on how the resources reflect the morphological richness of their languages (Section 2). Based on the survey's results, we suggest that moving away from the traditional bilingual dictionary structure and moving toward using computational models of the languages' rich morphological systems is a desirable direction for lexicography of low-resource, morphologically-complex languages, and is a "low-hanging fruit" in the application of language technology (Section 3), the pursuit of which has been passionately advocated and recommended by Trond Trosterud, and which has been a great inspiration for the authors.

2 The metalexicographical survey

Our goal in conducting this survey was to determine how online resources for Indigenous languages in Canada convey information about the complex morphology and inflectional paradigm-formation of these languages. This question is interesting because, as we mentioned above, resources whose data is stored in lexical databases are inherently hard-pressed to do justice to a morphologically rich language, and ethnographic linguists and lexicographers must come up with alternative strategies to convey the language's morphology.

For illustration, consider the internal complexity of the Plains Cree word *niwâpamâw* in (1) (from Harrigan et al. 2017: 570):

(1) ni-wâpam-â-w
1SG.SBJ.INDP-see.VTA-THM-1SG.SBJ.3SG.OBJ[1]
'I see him/her (animate)'

A lexical database resource essentially has two choices when representing a wordform like this. It could take the unit *niwâpamâw* as the headword and equate it with the phrase 'I see him/her', relying on the user to deconstruct the inflectional morphology, or it could take the root *wâpam-* 'see' as the headword and rely on the speaker to build up the surrounding morphology. In either case, the user needs further information about how to deal with the morphemes required to use this word accurately. How do lexical database resources for Indigenous languages in Canada convey this information? And what other lexicographical infrastructures exist that sidestep this issue entirely? These questions will be explored in the rest of this section.

[1] Abbreviations from Harrigan et al. (2017: 568): 1SG.SBJ: first person singular subject; INDP: Independent order; VTA: Verb Animate Transitive; THM: Theme sign; 3SG.OBJ: third person singular object.

2.1 Method

We began by taking the list of Indigenous languages recognized by Statistics Canada (2011) and searching in Google "[name of language] online dictionary".[2] In this way, we discovered the seventeen resources listed in Table 1.

These resources are all bilingual between the Indigenous language and English (and in a few cases, Canada's other official language, French). They range from simple HTML/XML-structured word lists to intricate websites that rival majority languages online dictionaries. Three of the websites – First Voices, Mother Tongues Dictionaries, and Ohwejagekha: Ha'degaenage – are multi-database interfaces that contain data from many Indigenous languages. This reduces the overhead of maintaining a website, making it simple for a community initiative to publish their language data. In the survey, we included the largest resource from each of these collections.

Language(s)	**Resource name**
Algonquin	The Algonquin Way Dictionary
Dakota	Dakota Dictionary Online
East Cree	Eastern James Bay Cree Dictionary
Gitksan	Gitksan/English Online Dictionary (Beta)
Hílzaqv	Mother Tongues Dictionaries: Hílzaqv
Mi'kmaq	Mi'kmaq Online
Mohawk	Ohwejagekha: Ha'degaenage: Mohawk
Moose Cree, Swampy Cree	Spoken Cree
Northern Sta̓ta̓timcets	First Voices: Northern Sta̓ta̓timcets
Ojibwe	The Ojibwe People's Dictionary
Passamaquoddy-Maliseet	Passmaquoddy-Maliseet Dictionary
Plains Cree	itwêwina
Plains Cree	Online Cree Dictionary
SENĆOŦEN	SENĆOŦEN Classified Word List
Sm'algyax	Sm'algyax Living Legacy Talking Dictionary
Tłı̨chǫ Yatıì	Tłı̨chǫ Yatıì Multimedia Dictionary
Tlingit	Online Tlingit Verb Dictionary

Table 1. The online resources for Canadian Indigenous languages included in the 2019 survey.

We originally evaluated each dictionary according to a broad set of criteria about how electronic lexicography can improve on the print dictionary tradition (primarily developed from Lew 2012, Granger 2012, and Prinsloo 2012). Many of these criteria exceed the scope of the current paper. For example, we evaluated whether the dictionaries incorporate multimedia elements and what sorts of extralexicographical (e.g. cultural, pedagogical) information is also available, and we recorded metrics like how many clicks it takes to get from the resource's front page to a page with information about the desired word (see Lachler and Pankratz 2017 for some of these results). Concerning our current goal, the portrayal of morphological structure, the primary desideratum from the literature was that online resources take advantage of technology to incorporate "more and better information" (Granger 2012: 2). In the next section, we will discuss the wide range in how more and better morphological information is represented in online resources for Indigenous languages in Canada.

[2] This method will only deliver web pages that have been indexed by Google, so some smaller and less-accessed resources may have eluded us for this reason. However, we assume that the websites that are indexed are the ones oriented toward interested users who might search for information about the language, and these are the resources that we want to explore.

2.2 Results: The emergence of two lexicographical approaches

In the seventeen resources we surveyed, we observed three main strategies for conveying morphological information, summarized in Table 2. The resources can be grouped according to which of these (sub)strategies they use. The number of resources using each strategy is shown in Table 2, and in Table 3, we have grouped the resources according to the strategies from Table 2 that they employ.

The surveyed resources have two possible structures, as shown at the right of Table 2: sixteen resources appear to be built around lexical databases, and one uses a computational model of the language's morphology (which is nevertheless based on a lexical database). These two approaches will be discussed more deeply in Section 3, but for now, we will go into more detail about the three groups of strategies. A common thread connecting most of these resources is that individual orthographic word-forms have their own page on the website, which will be referred to as the "lexical entry" or simply "entry" for that word.[3]

	Strategy	Number	Structure
1.	The resource contains no morphological information.	4	
2.	Each inflected wordform has its own entry.	2	lexical database
3.	Each paradigm or root has its own entry.	11	
	3a. Morphemes also have their own entry.	1	
	3b. Multiple complex wordforms provided, no gloss.	2	
	3c. Multiple complex wordforms provided, glossed.	3	
	3d. Generalized morphological slot templates provided.	1	
	3e. Individual morphological slot templates in each entry.	1	
	3f. Generalized full paradigms provided.	2	
	3g. Individual full individual paradigms in each entry.	1	computational model

Table 2. The strategies for portraying morphological information and the number of resources we surveyed that used each of them.

Group	Resource name
1	Gitksan/English Online Dictionary (Beta) Ohwejagekha: Ha'degaenage: Mohawk Online Cree Dictionary The Algonquin Way Dictionary
2	SENĆOŦEN Classified Word List Tłı̨chǫ Yatıì Multimedia Dictionary
3a	First Voices: Northern Sta̓timcets
3b	Dakota Dictionary Online (public version)[4] Spoken Cree
3c	Mother Tongues Dictionaries: Híłzaqv Mi'kmaq Online The Ojibwe People's Dictionary
3d	Sm'algyax Living Legacy Talking Dictionary
3e	Online Tlingit Verb Dictionary
3f	Eastern James Bay Cree Dictionary Passamaquoddy-Maliseet Dictionary
3g	itwêwina

Table 3. The surveyed resources listed according to how they express the morphology in their language (group numbers from the strategies listed in Table 2).

[3] The term "dictionary entry" often used for the same purpose, but since we are looking at resources that might not be considered dictionaries in the strictest sense, we have opted for "lexical entry" here.

[4] Lemma linking based on linguistic analysis is available for registered users of this resource, but we have evaluated the freely, publicly available version for the current survey.

2.2.1 Group 1: The resource contains no morphological information

Resources were grouped together here if only a headword in the Indigenous language and an English or French translation are provided, with no information about any morphological structure. As Table 3 shows, four of the seventeen surveyed resources fall into this group. That said, two of these only advertise themselves as word lists or phrase lists, so holding them to a higher lexicographical standard may be unfair. In any case, such lists are the bones of a lexicographical resource and a valuable first step in language documentation (cf. Bowern 2015).

2.2.2 Group 2: Each inflected wordform has its own entry

The two resources belonging to this group contain morphological information to the extent that the resource does contain multiple documented wordforms from the same lemma and sometimes an entire paradigm's worth of wordforms, but they are not conceptually grouped together. Rather, they are all listed at the same level as everything else, each as a single entry in the resource. This is a step forward over Group 1 because the resource reflects some of the wordforms possible in the language, though there is room for more abstraction away from the surface and toward the underlying paradigms.

2.1.3 Group 3: Each paradigm or root has its own entry

Like Group 2 resources, the ones in this group successfully include several wordforms, but beyond Group 2, these resources abstract away from the multiple orthographic words to their belonging to the same lemma/lexeme, which is what ties them all together. This step of abstraction is performed to different degrees of extensiveness and generalization, which is why this group has been divided into seven subgroups.

Group 3a resources recognize the morphemes as separate entities and list them parallel to the stems in the database (in one case, without information about how stems and morphemes can be combined to create new wordforms). The next two groups, 3b and 3c, have in common that a lexical entry may contain other wordforms which are morphologically related to the headword. What differentiates the groups is that in Group 3b resources, these wordforms are given without glossing or translation, while Group 3c resources also explain the morphological characteristics of the additional wordforms. For example, in Group 3c a noun's lexical entry might list another form as the plural, or a verb's lexical entry might provide and gloss that verb's first person singular form.

What Groups 3d and 3e share is one more step toward abstraction, namely that in both subgroups, morphological templates are provided (i.e. abstract schemas listing the order in which morphemes and stems appear in a morphologically-complex word). The templates in Group 3d resources are generalized versions which are valid for particular inflection classes. The templates given in Group 3e resources are not applicable to wider classes of stems, but given individually in each lexical entry, with the headword already integrated into the template itself.

At this point, as long as information about the morphemes themselves is provided, we begin to arrive at the resources which make it possible to construct a correctly-inflected word only from the information in the resource itself. However, this morphological information is not always there. When it is available, it is often in the form of lists of prefixes, verbal themes, etc., requiring quite in-depth linguistic knowledge to figure out how it should all be combined. All in all, the Group 3d and 3e resources we surveyed are linguistically very detailed, but might not be as accessible to community members some without linguistic training.

Finally, the last two groups are those which combine linguistic informativeness with layperson-oriented accessibility. Groups 3f and 3g include those resources which offer full inflectional paradigms (and enough information that the user can quickly find the paradigm for the word class they are interested in). The division between these is parallel to the division between the two groups of template-based resources above. Group 3f provides generalized paradigms which are representative of a particular word class, while every lexical entry in a Group 3g resource contains its own individual full paradigm with all core inflectional forms (though more derivational affixation have been left out).

As Table 2 shows, Group 3f is the last group where we observe a resource constructed around a lexical database. With this type of structure, it would not be easy to include full paradigms for each individual root contained in the resource, especially because so much of the stored information would be redundant, appearing over and over again in different but related paradigms. This is where the computational model of the language's morphology found in Group 3g shows its strengths and value for providing quick access to detailed linguistic information.

2.2.4 The take-away

In sum, the traditional conceptual layout of a bilingual dictionary – linking a single headword in one language to an equivalent in the other – is maintained in most of the resources we surveyed. This is unsurprising, since online language resources built around lexical databases are the *de facto* standard (cf. Prinsloo 2012; Lew 2012; Granger 2012), and not without reason. Descriptive linguists often begin by collecting lists of inflected wordforms and phrases before delving more deeply into the morphology and subtler linguistic structures (cf. Bowern 2015). The software that these linguists use – Toolbox and the like – digitizes these word lists in lexical databases, enforcing the one-to-one mapping between words in each language.

Even though this database structure is limited in how it can express morphological richness, we have observed certain strategies by which the surveyed resources nevertheless convey morphological information.

In what follows, we do not wish to downplay the hard work that goes into making any lexicographical resource, whether built around a lexical database or a computational model. However, we would like to argue that one of the most promising benefits that the digital domain offers to language resources is that of complete integration, especially for low-resource or endangered languages (cf. Prinsloo 2012: 127), and that this direction should be pursued further in the lexicography of Indigenous languages in Canada and elsewhere.

3 Discussion: Reinterpreting the Boasian trilogy

Our metalexicographical survey has shown that language resources constructed around a lexical database do their best to overcome the limitations of this format by employing strategies like providing morphological templates or linking the user to relevant paradigms containing the core inflectional forms that the user may want. These are good measures for creating valuable linguistic resources. However, any lexical database resource keeps the grammar outside of the core of the resource by design, no matter how detailed the morphological information is that it provides. Here, we would like to follow Prinsloo (2012) and argue that a digital lexicographical resource should be an all-in-one, integrated tool with all the lexicographic information that any user group should need. This is a demanding goal that cannot be achieved without rethinking lexicographical and documentary traditions and trying something new, adapting long-held methodological cornerstones to meet modern needs.

One such cornerstone in documentary linguistics is the production of a grammar, a dictionary, and texts in a language: the 'Boasian trilogy' (cf. Rice 2011: 192–193; Woodbury 2011: 163), so called based on Boas' (1917: 1) observation that in Indigenous language research "[w]e have vocabularies; but, excepting the old missionary grammars, there is very little systematic work. Even when we have grammars, we have no bodies of aboriginal texts."

Having this set of resources is valuable not just for linguistic research, but also for language documentation and revitalization in the community. The very name makes it clear that there are three distinct parts, but the three-way partition can be integrated into one resource using digital technology. We saw this in Group 3g in our survey, which contains a resource for Plains Cree named *itwêwina* (literally the plural noun form 'words' in Plains Cree). We only discuss the morphological integration aspect of *itwêwina* here, but it is indeed outfitted with a corpus of Plains Cree texts as well, completing the 'trilogy'. We must also concede here that we are leaving out in this discussion syntactic knowledge as an important part of grammar, though one has to recognize that in such morphologically rich and complex languages the

morphosyntactic component, i.e. inflected wordforms, is the one undertaking much of the "heavy lifting" in grammar.

itwêwina (https://itwewina.altlab.app) is built around a set of finite state transducers (FSTs, cf. Beesley and Karttunen 2003) which map between the orthographic wordform and its abstract morphological and morphosyntactic features and root. This process works in either direction, either encoding (producing an inflected form of a lemma based on morphological and morphosyntactic features) or decoding (parsing an already-inflected word; cf. Johnson et al. 2013: 62 for North and South; Harrigan et al. 2017 for Plains Cree verbs; Snoek et al. 2014 for Plains Cree nouns). In an FST-based online dictionary, the user can search with a fully-inflected orthographic word, and the computational model will deconstruct it into its stem and affixes, provide information about the morphemes, and direct the user to the appropriate entry in the dictionary. (This may be possible for lexical database resources too, but none of the ones we surveyed had this capability.)

Resources like *itwêwina* that can morphologically analyze and decompose complex words are a step toward a new state of the art in digital lexicography of Indigenous languages. Other languages in the world have resources which are also making this step. Turning to the Bantu languages of Africa, we find the website *isiZulu.net*, the most sophisticated resource available for a Bantu language (Prinsloo 2012: 135). Like *itwêwina*, the user can enter a fully-inflected orthographic word in the search bar, and the website returns the dictionary entry for the root of that word, as well as information about the original word both as linguistic glosses and as plain-speech translations, which makes the resource viable for communities without specialist linguistic knowledge. In the United States, Arapaho also has a resource built around finite state machines (Kazeminejad et al. 2017). And the same FST framework used by *itwêwina*, *Neahttadigisánit* (literally 'internet digital words' in North Saami), was first developed for the Saami languages of Northern Scandinavia by Giellatekno at UiT Arctic University of Norway (Tromsø), founded and directed by Trond Trosterud (cf. Johnson et al. 2013), and then brought to Canada for use with the morphologically complex languages there. *itwêwina* is the best-developed resource with this technology in Canada, but comparable resources for Northern Haida, Tsuut'ina, East Cree, and Odawa are currently in development (cf. Lachler et al. 2018; Arppe et al. 2017a; Arppe et al. 2017b; Bowers et al. 2017).

Indeed, this extensibility is one of the advantages of such a computational model. Adapting an existing model to new but typologically related languages is fairly straight-forward (if the linguistic documentation is there), much easier than compiling an entirely new lexical database from scratch, offering much smaller languages the same chance at excellent linguistic resources.

Another advantage of computational-model-based dictionaries is how much of the language they can encompass; they are arguably "the only way of ensuring good coverage of the language" (Trosterud 2004: 92). This is because adding one new root into an FST-based system results in the ability to decode any number of inflected wordforms built around that root. In contrast, adding one word into a lexical database dictionary results in the ability to decode exactly that one word more (Johnson et al. 2013: 69). In languages like Plains Cree and others in the Algonquian family, where the number of forms in a paradigm is in principle infinite (because one can add as many preverbs as one likes, though even without these, the paradigms of certain verb classes can be hundreds large), it is simply not feasible to try to contain all or even most words in a lexical database by adding them one by one. Assembling the computational morphological model is a great deal of work at the beginning, though surprisingly straight-forward provided that the linguistic description has been meticulous in determining morphological behavior of words, but it pays off immensely over time. If such extensive and comprehensive linguistic description already exists, as was the case for Plains Cree with the lexical database underlying the bilingual Cree-English dictionary, *nêhiyawêwin : itwêwina / Cree : Words*, diligently compiled by Wolvengrey (2001), then developing a computational model based on such a database and next expanding the database with that model turned out to be a veritable "low-hanging fruit", as Trond Trosterud has pitched it, in the application of language technology.

Above, we encountered the issue that sophisticated linguistic knowledge would sometimes be required to build correctly- inflected wordforms based only on the information given in the resource. This is another difficulty that is overcome by the computational model, because it is possible to display the information from one underlying model in many different ways. For example, *itwêwina* has linguistic descriptions of

the paradigms ("first person singular actor → third person singular goal", "Present tense"), as well as layperson-oriented ones in plain English (e.g. "I → him/her", "Something is happening now"), and even in "plain" nêhiyawêwin, i.e. Cree (e.g. "niya → kiya", "ê-ispayik anohc/mêkwâc/mâna"). This helps overcome the linguistic terminology barrier and makes the resource much more accessible to the community and non-specialists in general.

Another benefit of generating full paradigms is that the language user in search of a word expressing particular morphological characteristics on a particular stem will be able to find one, even if it has not been attested in the existing documentary corpus for that language. This allows the user to express the specific meaning they intend in much the same way that native speakers do, relying on the existing morphological patterns in the language.

Also, having such a model means that one can use it not just locally in an online dictionary, but in other tools as well. For example, Plains Cree also has a spell-checker under development (cf. Arppe et al. 2016: 5–6), a computer-aided language learning resource called *nêhiyawêtân* (literally 'Let's speak Cree'; cf. Bontogon et al. 2018), and a reader plugin, which allows the user to click on a word in Plains Cree appearing on any website and see a pop-up providing the lemma resulting from the linguistic analysis of the word and the English translation of the lemma (cf. Johnson et al. 2013: 65).

Arguably one of the most crucial advantages of such a resource, especially for communities working on language revitalization, is its usefulness to learners. Encoding and decoding morphologically complex wordforms requires extensive morphophonological knowledge, which a learner is unlikely to have, at least at the beginning. A computational model that can encode and decode complex wordforms does the worst of the work for the learner, so they have more energy left over to learn. As time goes on and the motivated learner's competence increases, they will need to consult the resource less and less, so it is particularly valuable for the early stages of language acquisition that could otherwise be overwhelming.

Finally, we could imagine that Table 2's list of strategies for conveying morphological information extends even further into more complex computational domains like deep/machine learning applied on (large) text collections. However, we would argue that the FST-based method actually has an advantage over deep learning approaches, because it does not need as much data to create a successful model. As conventional wisdom has it, the amount of data necessary for tried-and-true computational learning methods would be at least ten million words.[5] This amount of data is very unlikely to be available for low-resource languages, but an FST can be created without a substantial corpus. Therefore, FSTs have been practically the only way to create an extensive computational resource for these languages (Trosterud 2004: 92). Intriguingly, it has recently been shown that a neural encoder-decoder model can learn from comprehensive collections of morphological paradigms to mimic the behaviour of a handmade FST; cf. Moeller et al. (2018).

All in all, there are many lexicographical, documentary, and pedagogical advantages to having a computational model of a low-resource language's morphology. However, some issues must also be considered before adopting this type of model, in particular the question of how important it is that all wordforms contained in the resource be attested, verified forms in the language. We will focus on this issue for the remainder of this section.

3.1 Documentary and computational considerations

A generative morphological tool like the one incorporated within *itwêwina* raises a principled question about generalization and overgeneralization of paradigms: Is it better for a lexical resource to represent only words that have been verified, upholding native speakers' intuitions about possibilities and restrictions in their language, or to cover as much of the language as possible in an urgent language endangerment situation?

We may understand these two positions in parallel to the two lexicographical approaches outlined above. Lexical databases contain those words which are attested through fieldwork or consultation with the language community, while a computational model produces maximum possible coverage of the language,

[5] Mans Hulden, personal communication, for the case of German morphology.

even at the cost of overgenerating forms that speakers may not actually use, which has been described as the assumption of "lexical generality" (Bybee 1985: 84–87; Arppe et al. 2000: 5). And different sorts of linguists will prefer one approach or the other: A primarily computational linguist accepts the possibility of mistakes or overgeneralized forms creeping into the resource, while a primarily documentary linguist prefers to check and test everything, letting only gold-standard data through, although it only represents a fraction of the language.

Different communities (and subgroups within these communities) will also likely react to this question very differently. In our experience, language communities at the extremes of endangerment – those with very few elders or native speakers left, *and* those with millions of speakers – both tend to be less bothered by possible issues of morphological overgeneration. For very small endangered languages, keeping the language generally alive is the greatest priority, even if some words that nobody actually uses arise in the process, and for majority languages, native speakers and fluent learners know how the language is used anyway, so a few extra forms are not an issue. Crucially, even the potentially overgenerated word-forms are created according to common morphological processes and regular word-formation strategies observed in the language in question, and allow for as close an approximation as possible as to what a particular form could be, even if it has not been attested (rather than not being able to utter that form, or have to use a periphrastic construction, or need to use a majority language, to express some set of linguistic features).

However, toward the middle of the endangerment spectrum, priorities may be different and reactions may be more mixed. Some communities or community members might object to the fact that not every word has been checked, while others might appreciate how much of the language such a model can generate. And other community factors also come into play, such as the total number of speakers, their degree of language proficiency, whether a standardized form of the language exists, what kind of language education is available, and so on (cf. Arppe et al. 2016: 3). The decision about which of the two drawbacks – an overproductive resource or not enough coverage of the language – is the lesser of two evils should be done together with the community and in consideration of the other factors at play.

Overall, the two approaches must inform one another. A purely computational resource without actual verified data, for example, is no good to anybody. But we believe that the computational approach is particularly useful for time-sensitive endangered language situations, especially with languages like those in Canada, where it is prohibitively infeasible to check every form of every paradigm. Do we wait until we have a completely finished, perfected model which exactly reflects speaker intuitions, and until then have nothing (never mind that no lexicographical projects are ever really finished)? Or do we accept that it will be imperfect and will make mistakes around idiosyncrasies and other areas of variation? We have argued in this paper for the value of the second approach, especially given its flexibility: a computational model can generate all possible wordforms and then we can mark out those that have been attested, to distinguish them from the ones that are purely the creations of the model. This validation can be done steadily over time to increase the set of attested wordforms, allowing the two methods to meet in the middle. This approach lets us work in a time-sensitive way and simultaneously do justice to the morphological complexity of Indigenous languages in Canada.

4 Conclusion

In our survey of seventeen online resources for Indigenous languages in Canada, we found that these resources pursue different strategies to convey their language's complex morphology. These strategies are applied to get around the inherent limitations of resources built around a lexical database for expressing the complex morphological (sub-word) structures that are the norm in Canadian Indigenous languages. We have shown that one resource unites aspects of the grammar and the lexicon and more accurately reflects the language's morphology by incorporating a computational morphological model.

We have also argued that the computational approach represents a step forward in the evolution of lexicography, especially for low-resource languages and their communities. For these communities, the reinterpretation of the Boasian trilogy and the print dictionary tradition means the creation of an integrated resource, one that contains enough information that speakers and learners can use it to correctly construct

the complex words in their language. The ability to use the language correctly and fully, and to discover these forms easily, can be a significant boon to revitalization efforts.

Furthermore, we argue that a computational morphological model allows us as linguists to give more back to the community than do standard lexical database dictionaries (cf. also Lachler and Pankratz 2017). Such a model can be used to develop diverse language technologies like parsers, spell-checkers, and language learning applications, which are all ways to use technology to place the language back into the hands of the community and bring it back into day-to-day life. Language resources that sit on shelves (or hard-drives), safely archived, may help linguistic research, but do not help language revitalization in the communities who helped create the resources in the first place. The incorporation of computational models into digital lexicography for low-resource languages is a way to take decades of work that has shaped and redefined much of linguistic theory and turn it into something useful for both linguists and the community – this often really is a low-hanging fruit in the application of language technology, ripe for the picking.

In short, we cannot forget that language documentation and revitalization is a race against time, and tools like computational models that encompass the morphological richness of these languages help us help these communities, do it well, and do it quickly.

References

Arppe, Antti, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N. Moshagen. 2016. Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. *Proceedings of CCURL 2016 – Collaboration and Computing for Under-Resourced Languages.* 1–8.

Arppe, Antti, Mari Voipio, and Malene Würtz. 2000. Creating Inflecting Electronic Dictionaries. Edited by Carl-Erik Lindberg and Steffen Nordahl Lund. *Proceedings of the 17th Scandinavian Conference of Linguistics, Nyborg*. 20–22.

Arppe, Antti, Christopher Cox, Mans Hulden, Jordan Lachler, Sjur N. Moshagen, Miikka Silfverberg and Trond Trosterud. 2017a. Computational Modeling of Verbs in Dene Languages: The Case of Tsuut'ina. In Alessandro Jaker (ed.), *Working Papers in Athabaskan (Dene) Languages,* 51–69. Alaska Native Language Center Working Papers 13. Alaska Native Language Center, Fairbanks.

Arppe, Antti, Marie-Odile Junker, and Delasie Torkornoo. 2017b. Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection. *Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2).* 43–47. https://doi.org/10.18653/v1/W17-0108.

Beesley, Kenneth R., and Lauri Karttunen. 2003. *Finite state morphology.* Studies in Computational Linguistics. Center for the Study of Language and Information, Stanford, California.

Boas, Franz. 1917. Introductory. *International Journal of American Linguistics* 1(1): 1–8. https://doi.org/10.1086/463708.

Bontogon, Megan, Antti Arppe, Lene Antonsen, Dorothy Thunder, and Jordan Lachler. 2018. Intelligent Computer Assisted Language Learning (ICALL) for *nêhiyawêwin*: An In-Depth User-Experience Evaluation. *Canadian Modern Language Review* 74: 337–362. https://doi.org/10.3138/cmlr.4054.

Bowern, Claire. 2015. *Linguistic fieldwork: A practical guide.* Springer. https://doi.org/10.1057/9781137340801.

Bowers, Dustin, Antti Arppe, Jordan Lachler, Sjur N. Moshagen, and Trond Trosterud. 2017. A Morphological Parser for Odawa. *Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2).* 1–9. https://doi.org/10.18653/v1/W17-0101.

Bybee, Joan. 1985. *Morphology: A study of the relation between meaning and form.* (Vol. 9, Typological studies in language). John Benjamins, Amsterdam. https://doi.org/10.1075/tsl.9.

Granger, Sylvaine. 2012. Introduction: Electronic lexicography – from challenge to opportunity. In *Electronic Lexicography*, edited by Sylvaine Granger and Magali Paquot, 1–12. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199654864.003.0001.

Harrigan, Atticus G., Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology* 27: 565–598. https://doi.org/10.1007/s11525-017-9315-x.

Johnson, Ryan, Lene Antonsen, and Trond Trosterud. 2013. Using Finite State Transducers for Making Efficient Reading Comprehension Dictionaries. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013).* 59–71.

Kazeminejad, Ghazaleh, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages – the case of Arapaho. *Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages (ComputEL-2).* 10–18. https://doi.org/10.18653/v1/W17-0102.

Lachler, Jordan, and Elizabeth Pankratz. 2017. Moving toward value-added digital repatriation in lexicography for Indigenous languages in Canada. Edited by Nicholas Ostler, Vera Ferreira and Chris Moseley. *Proceedings of the 21st FEL Conference: Communities in Control: Learning tools and strategies for multilingual endangered language communities.* 107–114.

Lachler, Jordan, Lene Antonsen, Trond Trosterud, Sjur N. Moshagen, and Antti Arppe. 2018. Modeling Northern Haida Morphology. *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan, May 7–12, 2018, 2326–2330.

Lew, Robert. 2012. How Can We Make Electronic Dictionaries More Effective? In *Electronic Lexicography*, edited by Sylvaine Granger, and Magali Paquot, 343–361. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199654864.003.0016.

Moeller, Sarah, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for Arapaho verbs learned from a finite state transducer. *Proceedings of Workshop on Polysynthetic Languages.* Association for Computational Linguistics, 12–20. https://aclanthology.org/W18-4802.

Prinsloo, D. J. 2012. Electronic lexicography for lesser-resourced languages: The South African context. In *Electronic Lexicography*, edited by Sylvaine Granger, and Magali Paquot, 119–143. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199654864.003.0007.

Rice, Karen. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation* 5: 187–207.

Statistics Canada. 2011. Aboriginal languages in Canada. *2011 Census of Population, Catalogue no. 98-314-X2011003.* https://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011003_3-eng.pdf (accessed April 19, 2019).

Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. *Proceedings of ComputEL: Workshop on the use of computational methods in the study of endangered languages*, 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, 26 June 2014, 34–42. ACL Anthology. https://doi.org/10.3115/v1/W14-2205.

Spencer, Andrew. 2016. Two morphologies or one? Inflection versus word formation. In *The Cambridge Handbook of Morphology*, edited by Andrew Hippisley and Gregory Stump, 27–49. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781139814720.002.

Stump, Gregory. 2016. *Inflectional paradigms: Content and form at the syntax-morphology interface.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781316105290.

Trosterud, Trond. 2004. Porting morphological analysis and disambiguation to new languages. *First Steps in Language Documentation for Minority Languages: Proceedings of the SALTMIL Workshop at LREC 2004.* 90–92.

Trosterud, Trond. 2006. Grammatically based language technology for minority languages. *Lesser-Known languages of South Asia: Status and policies, case studies and applications of information technology.* Mouton de Gruyter, Berlin. 293–316. https://doi.org/10.1515/9783110197785.

Wolvengrey, Arok. 2001. *nêhiyawêwin: itwêwina / Cree: Words*, bilingual edition. University of Regina Press, Regina.

Woodbury, Anthony C. 2011. Language documentation. In *The Cambridge Handbook of Endangered Languages*, edited by Peter K. Austin and Julia Sallabank, 159–186. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511975981.009.

Language resource references

Eggleston, Keri. *Online Tlingit Verb Dictionary*. http://ankn.uaf.edu/~tlingitverbs/ (accessed April 19, 2019).
Ellis, Doug. *Spoken Cree: Moose and Swampy Cree Dictionary.* http://www.spokencree.org/Glossary (accessed April 19, 2019).
First Voices. *Northern Státimcets.* http://www.firstvoices.com/en/Northern-Statimcets (accessed April 8, 2019).
isiZulu.net. *isiZulu.net: Bilingual Zulu-English dictionary.* https://isizulu.net/ (accessed April 19, 2019).
Junker, Marie-Odile, Marguerite MacKenzie, Luci Bobbish-Salt, Alice Duff, Linda Visitor, Ruth Salt, Anna Blacksmith, Patricia Diamond, and Pearl Weistche. *The Eastern James Bay Cree Dictionary on the Web: English-Cree and Cree-English, French-Cree and Cree-French (Northern and Southern dialects).* http://dictionary.eastcree.org/words (accessed April 19, 2019).
Mi'kmaq Online. *Mi'kmaq Online.* https://www.mikmaqonline.org/ (accessed April 19, 2019).
Miyo Wahkohtowin Community Education Authority. *Online Cree Dictionary.* http://www.creedictionary.com/ (accessed April 19, 2019).
Mother Tongues Dictionaries. *Online Hiłzaqv Dictionary.* https://mothertongues.org/heiltsuk/ (accessed April 19, 2019).
Ohwejagekhá: Ha'degaenage. *Mohawk.* http://ohwejagehka.com/mohawk/ (accessed April 19, 2019).
Omàmiwininì Pimàdjwowin/The Algonquin Way Cultural Centre. *The Algonquin Way Dictionary.* http://www.thealgonquinway.ca/English/dictionary-e.php (accessed April 19, 2019).
Passamaquoddy-Maliseet Language Portal. *Passamaquoddy-Maliseet Dictionary.* https://pmportal.org/browse-dictionary (accessed April 19, 2019).
SENĆOŦEN Classified Word List. *SENĆOŦEN Classified Word List.* https://itservices.cas.unt.edu/~montler/Saanich/WordList/ (accessed April 19, 2019).
The University of Northern British Columbia. *Sm'algyax Living Legacy Talking Dictionary.* http://web.unbc.ca/~smalgyax/ (accessed April 19, 2019).
University of Alberta ALTLab. *itwêwina.* https://itwewina.altlab.app/ (accessed April 19, 2019).
University of British Columbia Department of Linguistics. *Gitksan/English Online Dictionary (Beta).* http://gitdict.nfshost.com (accessed April 19, 2019).
University of Minnesota Department of American Indian Studies. *Dakota Dictionary Online.* https://filemaker.cla.umn.edu/dakota/ (accessed April 19, 2019).
University of Minnesota Department of American Indian Studies. *The Ojibwe People's Dictionary.* https://ojibwe.lib.umn.edu/ (accessed April 19, 2019).
University of Victoria Linguistics Department. *Tłı̨chǫ Yatìi Multimedia Dictionary.* http://tlicho.ling.uvic.ca/ (accessed April 19, 2019).

Samiske barnehagers rolle i språkrevitaliseringa

Torkel Rasmussen
Samisk høgskole

Abstract

More than a thousand children attend Sámi kindergartens daily, while quite a number of Sámi children get Sámi language instruction in other kindergartens. This activity is one of the most important arenas for transmission and acquisition of Sámi languages. A question raised in this article is how these kindergartens are used as research fields in the disciplines of language and sociology and early childhood pedagogy. Another question is what kind of language teaching models are used. The article shows that little research has been carried out on how Sámi kindergartens teach Sámi language. There is also little research done on the results from this education. A review of the Sámi kindergartens' history shows that statistical material on Sámi kindergartens and Sámi language instruction in other kindergartens is available only from some geographical areas, and existing statistical information is only partly suitable for analysis. This makes it difficult to use existing material to monitor the vitality of Sámi languages. The article calls for more research on Sámi kindergartens and language teaching models used in them. A goal could be to create a basis for monitoring this crucial indicator of language vitality for Sámi languages: whether new generations of Sámi become Sámi-speakers or not. This should be followed up with more research on the language teaching models used in Sámi kindergartens.

Keywords: Sámi kindergarten, language nest, language revitalisation, language planning

1 Innledning

«Hvilken rolle spiller barnehagene i dag i arbeidet med å styrke stillinga for samiske språk?». Dette er spørsmålet som blir behandlet denne artikkelen som tar for seg barnehagetilbudet for samiske barn i førskolealder i Sápmi[1] – og i byer utenfor Sápmi.

Samiske barnehager fins, i et antall av 60–70, i svært mange samiske lokalsamfunn og i en god del nordiske byer. Den samiske barnehagen har en mer enn 50 år lang historie, og svært mange samer har etter hvert et forhold til den, enten som tidligere barnehagebarn eller som foreldre og besteforeldre til barnehagebarn. Til tross for dette har det vært forsket relativt lite på samiske barnehager, og vi har av den grunn lite forskningsbasert kunnskap om livet i barnehagen eller om resultatet av et opphold i en samisk barnehage i forhold til målet om å styrke barnas samiske språk, kultur og identitet.

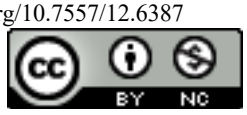
Dette gjør den samiske barnehagen til et spennende fagfelt der det er mulig å gjøre banebrytende nybrottsarbeid. Det er mulig å forske på samiske barnehager fra forskjellige perspektiver. Et lingvistisk perspektiv kan være å utforske barnespråk (Ijäs 2011), et språksosiologisk perspektiv kan se på barnehagens rammevilkårene og rolle i språkrevitaliseringa (Storjord 2004; 2006; 2008; Todal 2007; Rasmussen 2005; 2013a; 2013b; Laiti 2018; 2019), og et pedagogisk perspektiv kan legge vekt på barnas lek, samarbeid og samspill med de ansatte, og ikke minst på de metodene som blir brukt for å overføre samiske språk til den neste generasjonen samer (Storjord 2008; Braut 2010; Kleemann 2015; 2021; Pasanen 2015; Äärälä-Vihriälä 2016.)

Det er lett å se hvordan disse perspektivene kan gi mange interessante tverrfaglige tilnærminger, men i denne artikkelen blir fokuset hovedsakelig rettet mot to forhold:

1. Hvilke språkopplæringsmodeller brukes i barnehagene for å utvikle barns samiske språk?
2. Kan det tilgjengelige materialet om de samiske barnehagene brukes til å vurdere samiske språks vitalitet?

[1] Det samiske bosetningsområdet i Norge, Sverige, Finland og Russland.

 Nordlyd 46.1: 205–217, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigert av Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publisert ved UiT Norges arktiske universitet. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6387



Artikkelen består av seks kapitler. I kapittel 2 redegjøres det for noen sentrale begreper og for metodene som er brukt i artikkelen. I kapittel 3 presenteres noen sentrale forskningsbidrag om den samiske barnehagen. Kapittel 4 gjennomgår noen av de relevante språkopplæringsmodellene som brukes for å gjøre barn tospråklige, og er et forsøk på å utvikle kategorier for samiske språkopplæringsmodeller som brukes i barnehagen. Kapittel 5 tar for seg de samiske barnehagenes historie, nåværende utbredelse og undersøker hva vi vet om språkopplæringsmodellene som er i bruk i barnehagene. I kapittel 6 diskuteres funnene, mens jeg i konklusjonen peker på behovet for forskning på den samiske barnehagen og de språkopplæringsmodellene som brukes.

2 Begreper og metode

2.1 Definisjon av samiske barnehager

Det fins ikke en enhetlig definisjon på hva en samisk barnehage er. Sametinget i Norge (2022) har utviklet en definisjon som brukes for tildeling av økonomisk støtte til samiske barnehager og samiske barnehageavdelinger i norske barnehager. For å være støtteberettiget som samisk barnehage eller samiske barnehageavdelingene i Norge, må de ha vedtektsfestet at barnehagetilbudet bygger på samisk språk og kultur. Kriteriene for en samisk barnehage er at barnehagen ledes av samiskspråklig pedagogisk personale, ansatte i barnehagen er samiskspråklig og driftsspråket er samisk. Et kriterium for å bli godkjent som ei samisk barnehageavdeling, er at ansatte i avdelinga skal være samiskspråklige og driftsspråket skal være samisk. Den samiske barnehagen defineres derfor i denne artikkelen som:

> *En samisk barnehage eller barnehageavdeling der tilbudet bygger på samisk språk og kultur, de ansatte er samiskspråklige, driftsspråket er samisk og målet er å overføre samisk språk til barna.*

Marikaisa Laiti har drøftet definisjonsproblemet i sin doktorgradsavhandling i pedagogikk (Laiti 2018: 30). Hun undersøker det samiske førskoletilbudet i Finland og bruker begrepet S*aamelainen varhaiskasvatus* – (samisk småbarnspedagogikk) som et samlebegrep for de to språkopplæringsmodellene som brukes i Finland: 1. morsmålsbarnehager for samiskspråklige barn og 2. samiske språkreir. Det sistnevnte er et tilbud på samisk for barn som ikke behersker samisk når de begynner i barnehagen.

Det fins også samiske barnehager der tilbudet er mer innrettet mot samisk kultur og identitet enn språk. Dette gjelder bl.a. den samiske barnehageavdelinga i Luujäu′rr/Lovozero i Russland (se kap. 5.4). Denne avdelinga omtales også i artikkelen til tross for at den faller utenfor begge definisjonene av en samisk barnehageavdeling ovenfor.

2.2 Språkskifte, språkbevaring og vending av språkskifte

Den samiske språksosiologiske situasjonen er sterkt preget av språkskifter. I alle fire land der samer bor, har det skjedd språkskifte fra samiske språk til landets majoritetsspråk. Det mest typiske språkskiftet skjedde når samiskspråklige foreldre ikke overfører samisk til sine egne barn. Det er store lokale forskjeller i Sápmi for når språkskiftet startet og hvor omfattende det var. Alle samer tok ikke del i språkskiftet, og noen opprettholdt samisk språk ved å bruke det aktivt også overfor egne barn. En vending av språkskiftet er det motsatte av det som er beskrevet ovenfor. Etterkommerne av dem som snakket minoritetsspråket, snakker i utgangspunktet ikke minoritetsspråket, men lærer seg det og blir (minst) tospråklige. I vendinga av språkskiftet spiller barnehagene ofte en sentral rolle. (Fishman 1991; Todal 2002; 2007; Rasmussen 2013a; Pasanen 2010; 2015; 2018; Huss 1999; Janson 2005; Johansen 2009; Hinton og Hale 2001: kap. 11–13).

2.3 Metode

Forskingslitteratur om samiske barnehager og språkopplæringsmodeller er valgt utfra tidligere kjennskap til feltet, gjennom søk på Google Scholar, og ved å gjennomgå litteraturlistene i publikasjonene som ble funnet, for på denne måten å finne annen relevant litteratur. Forskningspublikasjonene har blitt nærlest for å finne informasjon om de språkopplæringsmodellene som er i bruk i samiske barnehager og for opplæring i samisk språk i andre barnehager.

Materialet til kapittel 5 "De samiske barnehagenes historikk" kommer fra en rekke kilder. Viktige bidrag til historien har kommet fra forskere innen pedagogikk (Storjord 2004: 2006; 2008; Laiti 2018; 2019; Äärälä-Vihriälä 2016). Andre viktige kilder er utredninger og sakspapirer fra Sametingene (Nuorgam 2019; Sametinget i Sverige 2011; 2013; 2021). For eksempel oppdaterer Sametinget i Finland årlig statistisk materiale om de samiske barnehagene i Finland med detaljerte opplysninger om sted, kommune, antall barn og språkopplæringsmodell (Sametinget i Finland 2020). Sametinget i Norge publiserer årlig en del informasjon om samiske barnehager, bl.a. antall samiske barnehager, samiske barnehageavdelinger i norske barnehager, antall barn i disse barnehagene og antall barn som får samisk språkopplæring i norske barnehager. (Sametinget i Norge 2021). Denne informasjonen publiseres også av Statistisk sentralbyrå (SSB 2021).

De tilsvarende tallene fra Sverige var svært vanskelig å innhente, og det lyktes heller ikke fullt ut. Det samiske barnehagetilbudet i Sverige er dels drevet av *Sameskolstyrelsen* og dels av kommunene. Det fins ikke noe offentlig organ som har oversikt over antall barn i samiske barnehager eller antall barn som får opplæring i samisk i andre barnehager. En del informasjon for Sameskolstyrelsens barnehager fins i Sametingets årlige *Lägesrapport: De samiska språken i Sverige* (Sametinget i Sverige 2013: 2021). Men rapportene inneholder bare sporadisk informasjon om kommunenes samiske barnehagetilbud og ikke noe om språkopplæring i andre svenske barnehager. Slik informasjon må innhentes fra hver enkelt kommune. Dette er en svakhet ved artikkelen som det redegjøres for i kapittel 5.

Materialet om situasjonen i Russland er samlet inn av avdøde Jevgenij Jushkov (2020) som intervjuet barnehageansatte i Russland om den samiske barnehagens historie der. Ansatte ved barnehagelærerutdanninga på Samisk høgskole har også bidratt med materiale om den samiske barnehagens historie i Norge og Sverige.[2] Det sistnevnte bestod av muntlig informasjon gjennom samtaler, noen skriftlige notater og forelesningspresentasjoner. (Kuhmunen 2019/2020 og Eriksen 2019/2020).

Det blir i denne artikkelen lagt vekt på et allsamisk perspektiv ved å gjøre rede for den samiske barnehagesituasjonen i alle de fire land som har lagt under seg deler av Sápmi: Norge, Sverige, Finland og Russland. Imidlertid vil ei inndeling bare på land skjule den interne inndelinga av samiske språk, og det er av den grunn lagt vekt på å legge fram empiri om samiske barnehagetilbud også på de samiske minoritetsspråkene.

3 Tidligere forskning på den samiske barnehagen

Den samiske barnehagen er et relativt lite utforsket emne. Forskerne som har engasjert seg i dette feltet, er en håndfull, og de har litt forskjellige innfallsvinkler til emnet. Den ene delen kommer fra pedagogikkforskninga og den andre fra språksosiologien. En av de første til å undersøke samiske barnehagebarns vilkår var Juha Guttorm som undersøkte språkbruken i tospråklige samisk/finske barnehager i Utsjoki kommune. Han fant ut at bruken av samisk i tospråklige samisk/finske barnehager var tilfeldig og marginal. Av den grunn foreslo han opprettelse av egne samiskspråklige barnehager. (Guttorm 1986: 30). Marianne Storjord (2004) har i sitt masterarbeid forsket på den samiske barnehagens historie i Norge. Hun viser at bruken av samisk språk i de første barnehagene var tilfeldig, spesielt i de kommunale barnehagene i Indre Finnmark. I doktorgradsstudiene sine (2008) undersøkte hun den sosiale praksisen i barnehagene, blant annet ved observasjon og intervju med foreldre og ansatte. Hun konkluderte med at norsk dominerer i samiske barnehager, mens de norsktalende barna falt utenfor når det ble brukt samisk. Hun advarte om at barnehagene ikke hadde konkrete modeller i arbeidet med tospråklighet, og at de så ut til å mangle metodikk og et pedagogisk opplegg for å møte disse utfordringene. Carola Kleeman (2015) har undersøkt samiske barnehagebarns kodeveksling mellom samisk og norsk når de er i barnehagen. Hun kom fram til at barna, som var tospråklige, brukte begge språkene sine i rollelek, og at språkpraksisen er styrt av rollelekens regler. Kleemann (2021) har også tatt i bruk begrepet *transspråking* som opplæringsmetode på grunnlag av observasjoner hun har gjort i en samisk barnehage der barna (foreløpig) snakket lite samisk, men forstod mye. De ansatte fikk barna til å snakke litt samisk ved selv å veksle mellom samisk og norsk i sin kommunikasjon med barna.

[2] Forfatteren takker Anne Ingebjørg Svineng Eriksen og Gudrun Kuhmunen ved Samisk høgskole for bidrag til artikkelen.

På finsk side har Rauni Äärälä-Vihriälä (2016) forsket på hvordan samiske språkreir fungerer. Forskninga hennes er en casestudie av hvordan et samisk språkreir fungerer, og basert på resultatene utviklet hun en modell for en språkpedagogikk tilpasset samisk kultur for bruk i samiske språkreir. Marikaisa Laiti (2018; 2019) har i sin doktorgradsavhandling undersøkt hvordan samiske barnehager blir implementert i Finland. Dette er en studie av hva myndighetene har bestemt gjennom lovverket og reglement, og hvordan dette blir gjennomført på lokalplan. Hennes funn viser at det er problematisk å gi et samisk barnehagetilbud bygget på samisk språk og kultur, og samtidig følge de rutinene som storsamfunnet forutsetter for barnehagedrift. Hun peker også på et problematisk forhold ved lovverket. Et samiskspråklig barnehagetilbud er en rettighet for samiske barn som snakker samisk. Men de samiske barna som ikke snakker samisk, har ikke rett til å få et barnehagetilbud som inneholder opplæring i samisk språk. (Laiti 2018: 32). Laiti jobber nå (2021–23) som prosjektleder for forskningsprosjektet *Sámi mánáid-gárde-pedagogihkka ođđa áiggis* (Samisk barnehagepedagogikk i ei ny tid) ved Samisk høgskole. Målet med prosjektet er å støtte og utvikle samifiseringa av de samiske barnehagene. (Samisk høgskole 2022). Fra svensk og russisk side er det ikke registrert noe forskning på samiske barnehager og språkbruk.

Språksosiologenes interesse for de samiske barnehagene har vært en litt annen enn pedagogenes. De undersøker ofte språkets vitalitet, og da er tilgang på institusjoner for utvikling og revitalisering av språkene viktige positive faktorer, mens en mangel på slike institusjoner blir sett på som negativt for språkets vitalitet. Emeritusprofessor Mikael Svonni var en av de første som så på samiske barnehager i et slikt språksosiologisk perspektiv da han sammen med Kenneth Hyltenstam og Christopher Stroud utviklet en modell for å vurdere vitaliteten til minoritetsspråk og brukte denne til å undersøke de samiske språkenes vitalitet i Sverige (Hyltenstam og Stroud 1991; Hyltenstam, Stroud og Svonni 1999). Emeritusprofessor Jon Todal (2007) fulgte som forsker, et sørsamisk revitaliseringsprosjekt i *Svahken Sijte* der ungene ved bruk av en delvis språkreirsmodell lærte sørsamisk i barnehagen. Forfatteren av denne artikkelen har sin faglige bakgrunn fra språksosiologi der rammevilkårene for minoritet- og urfolksspråk er sentrale elementer. Jeg har undersøkt tilgangen til samiske barnehager både i mitt masterarbeid og min doktorgradsavhandling. Kvalitative intervju med foreldre til samiskspråklige barn viste at de var svært fornøyde med det samiske barnehagetilbudet, mens foreldre til ikke-samiskspråklige barn var svært misfornøyde med den samiskopplæringa deres barn hadde fått i barnehagen. (Rasmussen 2005; 2013a). I tillegg har jeg undersøkt de økonomiske rammene for de samiske barnehagene i Norge (Rasmussen 2013b). Professor i samisk språksosiologi ved Samisk høgskole, Annika Pasanen, har gjort flere undersøkelser av de enaresamiske språkreirene og hun har bl.a. undersøkt hvordan språkopplæringa skjer (2015; 2018). Disse forholdene blir beskrevet i mindre detalj av Leena Huss (1999) og Annika Jansson (2005).

Kristine Tjåland Braut (2010) undersøkte i sin masteroppgave språkbruken i en lulesamisk barnehage. Hun beskriver hvordan den gikk fra å være en morsmålsbarnehage for lulesamiskspråklige barn til å bli et lulesamisk språkreir fordi språkkunnskapene til foreldrene og barna endret seg over noen år. Hun observerte at barna snakker lulesamisk i barnehagen, og retter fokuset mot språkbruken utenfor barnehagen når hun diskuterer om barna kan bli funksjonelt tospråklige hvis de ikke får mer støtte fra foreldrene og andre utenfor språkreiret.

Både den pedagogiske og språksosiologiske tilnærminga til de samiske barnehagene har gitt oss ny kunnskap om språkopplæringsmodellene som blir brukt i de samiske barnehagene. Dette kan brukes til å sammenligne de samiske språkopplæringsmodellene med modeller som blir brukt i andre land og for andre (ur)folk.

4 Språkopplæringsmodeller

Det er et gjentakende problem i samisk forskning at man ikke kan gi eksakte tall for hvor mange samer det er. Estimatene for antall talere av de ti samiske språkene (Sammallahti 1998) spriker også, og det eneste en med sikkerhet kan si, er at det er flere samer enn talere av samiske språk (Øzerk 2006: 96). Dette fører også til at det i aldersgruppa 0–6 år vil være en del barn som lærer samisk hjemme og får tilbud om plass i en samisk barnehage. En stor del, kanskje et flertall av de samiske barna, lærer ikke samisk hjemme. Noen av dem får likevel plass i en samisk barnehage og mulighet til å lære samisk der, mens andre ikke får den muligheten. Noen av de samiske barna får tilbud om opplæring i samisk språk i norske og svenske

barnehager, mens andre samiske barn ikke får det. Bare fra Finland har vi gode tall på samiske barn og deres tilgang til et samisk barnehagetilbud fordi Sametinget i Finland har undersøkt dette forholdet. 82 prosent av de samiske barna i *Samernas hembygdsområde*[3] har et samisk barnehagetilbud, mens bare 7,5 prosent av de samiske barna utenfor dette området har et slikt barnehagetilbud (Nuorgam 2019: 8).

Faglitteraturen innen språksosiologi legger stor vekt på undervisningsspråket i skolen for den lingvistiske vitaliteten. Den framhever at det må være mulig å få undervisning enten på minoritetsspråket eller i det minste få tilbud om opplæring i minoritetsspråket som et skolefag. (Se Hyltenstam og Stroud 1991; Hyltenstam, Stroud og Svonni 1999; Fishman 1991 Baker 2001; UNESCO 2003). Disse forskerne har mindre søkelys på barn i barnehagealder, og det var nærmest et gjennombrudd innen forskningsfeltet da Kendall King (2009: 13–15) utfordret språksosiologenes syn på at barn er hjemme med mor til de begynner på skolen. I virkeligheten, påpeker hun, tilbringer barna mye tid utenfor hjemmet allerede i førskolealder og mindre tid hjemme med foreldrene.

I Sápmi *skal* barn få et samiskspråklig tilbud i barnehagen[4], og det er et offentlig ansvar å tilby foreldrene et tilbud som passer til deres barns språkferdigheter, og til lokalsamfunnets og foreldrenes språklige ambisjoner for barna. For at dette skal fungere, trengs det gode språkopplæringsmodeller.

En nestor innen tospråklighetspedagogikk internasjonalt er professor Colin Baker ved Bangor University i Wales. Han har forsket på tospråklige undervisningsmodeller som er i bruk forskjellige steder i verden i grunnskoler og videregående skoler. Han lanserte ideen om sterke og svake språkopplæringsmodeller i første utgave av boka *Foundation of Bilingual Education and Bilingualism* i 1993 og har utviklet modellen fram til syvende utgave i 2017. Hans viktigste skille er mellom sterke og svake modeller for tospråklig opplæring. De sterke tospråklige opplæringsmodellene har funksjonell tospråklighet som resultat. Barna behersker etter hvert både minoritets- og majoritetsspråket godt muntlig og skriftlig. Resultatet av de svake tospråklige undervisningsmodellene er enspråklighet eller begrenset tospråklighet. Resultatet er normalt at barna bare kan majoritetsspråket eller at de er bedre i majoritetsspråket enn i minoritetsspråket.

Baker (2001: 194) klassifiserer språkopplæringsmodellene som svake og sterke ut fra hvor mye de bruker minoritetsspråket i undervisninga. De svake modellene bruker majoritetsspråket som hovedspråk og enten bruker de ikke minoritetsspråket i det hele tatt, de tilbyr minoritetsspråket som et skolefag noen timer i uka, eller de bruker minoritetsspråket en del i begynnelsen av skolegangen og mindre og mindre etter hvert. Basert på forskningsresultater konkluderer han med at disse svake modellene gir enspråklighet i majoritetsspråket eller begrenset tospråklighet der elevene kan majoritetsspråket best.

De sterke språkopplæringsmodellene er: 1. språkbadsmodeller der majoritetsspråklige lærer minoritetsspråket fordi det er hovedspråk i undervisninga, 2. språkbevaringsmodellen[5] der minoritetsspråket er hovedspråk hele skoletida for minoritetsspråklige elever, og majoritetsspråket er et skolefag. I tillegg nevner Baker to typer tospråklig undervisning: En der både minoritetsspråket og majoritetsspråket er undervisningsspråk, og en annen der to majoritetsspråk er undervisningsspråk. Baker setter valg av språkopplæringsmodell i sammenheng med samfunnsmessige og pedagogiske mål. Han hevder at ettersom man vet at resultatet av svake modeller er enspråklighet eller begrenset tospråklighet, så er samfunnets mål stort sett assimilasjon av minoriteten. På den andre siden viser målene med de sterke språkopplæringsmodellene seg å være bevaring, revitalisering eller integrering fordi resultatet av språkopplæringsmodellen viser at elevene blir tospråklige.

I tabell 1 er noen av Bakers kategorier for opplæringsmodeller i grunnskolen og den videregående opplæringa brukt. Dette gjelder samfunnsmessig og pedagogiske mål og sannsynlig språklig utkomme. For øvrig er tabellen tilpasset de språkopplæringsmetodene empirien viser blir brukt overfor samiske barnehagebarn. Den er også tilpasset det vi vet om barnegruppene og om bruk av samisk i barnehagene ut fra

[3] I Finland kommunene: Enontekiö, Inari, Utsjoki og den nordlige delen av Sodankylä.

[4] Dette er lovregulert i alle fire land med lokale forskjeller som det ikke blir gått nærmere inn på her av plasshensyn.

[5] Av Øzerk (2006: 87) kalt den samiske førstespråksmodellen.

gjennomgangen av forskningslitteraturen, spesielt de eksemplene Øzerk (2006) redegjør for fra Sápmi. Tabellen må ikke leses som en ferdig konklusjon, men som et forslag til faglig diskusjon.

Språkopplæringsmodell	**Barnegruppe**	**Språk i barnehagen**	**Samfunns-messig og pedagogisk mål**	**Sannsynlig språklig utkomme**
Førstespråksmodellen	Samiskspråklige barn	Samisk	Bevaring	Godt samisk språk
Blandingsmodellen	Blandet gruppe: Samiskspråklige og ikke-samiskspråklige barn	Samisk	Bevaring, revitalisering/ integrering	Varierende resultat
Språkreirsmodellen	Ikke-samisk-språklige barn	Samisk	Revitalisering	Godt samisk språk
Utskillingsmodellen	Blandet gruppe: Samiske og andre barn. Få eller ingen samiskspråklige	Nasjonalstats-språk. Samiskopplæring for noen barn.	Fortsatt assimilering	Begrenset samisk-språklighet
Majoritetsspråksmodellen	Blandet gruppe: Samiske og andre barn. Få eller ingen samiskspråklige	Nasjonalstats-språk. Ingen samisk-opplæring	Fortsatt assimilering	Ingen eller begrenset samiskspråklighet

Tabell 1. Språkopplæringsmodeller i den samiske barnehagen, kategorisert etter barnegrupper, språk i barnehagen, samfunnsmessig og pedagogisk mål og sannsynlig språklig utkomme. Etter Colin Baker (2001: 194).

De tre øverste kategoriene i tabellen benytter sterke samiske språkopplæringsmodeller, og har bevaring, revitalisering og/eller integrering som samfunnsmessig og pedagogisk mål. Modell nr. 1 er førstespråksmodellen (av Baker kalt bevaringsmodellen). Den viser til samiskspråklige barnehager for samiskspråklige barn. Modell nr. 2 kalles blandingsmodellen fordi barnegruppa er ei blanding av samisk-språklige og ikke-samiskspråklige barn. Denne modellen er brukt i Norge og Sverige. Det er i liten grad forsket på om samiskspråklige barn får utviklet sitt samiske språk i ei slik gruppe, og om de i utgangspunktet ikke-samiskspråklige barna, lærer seg samisk godt i løpet av et opphold i en slik barnehage. I en samisk sammenheng må en kunne ha en hypotese om at det er svært store lokale forskjeller alt ettersom hvor stor andelen av ikke-samiskspråklige barn er i ei slik gruppe, hvor sterkt samisk språk står i lokalsamfunnet og hvor god kompetanse de ansatte i barnehagen har. Storjord (2008) viser varierende resultat som språklig utkomme.

Modell nr. 3, språkreirsmodellen, er et samiskspråklig barnehagetilbud for ikke-samiskspråklige barn. Den er brukt fra 1990-tallet i Finland. Denne metoden ble utviklet av maoriene på Ny-Zealand, og den ble også tatt tidlig i bruk for det beslektede urfolksspråket på Hawaii. (Hinton og Hale 2001: kap. 11–13). Ideelt sett lærer barna å snakke samisk i språkreiret fordi de voksne og barn som alt har lært samisk, først begynner å snakke språket til dem, og så med dem. Det er påvist gode resultater med de enaresamiske språkreirene i Finland (Jansson 2005: 124–127; Pasanen 2010; 2015; 2018). De skoltesamiske språkreirene kan foreløpig ikke vise til så gode resultater (Laihi 2017), og de nordsamiske språkreirene i Finland er i liten grad undersøkt.

De to siste språkopplæringsmodellene er svake samiskspråklige språkopplæringsmodeller. Den ene, kalt utskillingsmodellen (fra eng. submersion), tilbyr samiske barn et par timer samiskopplæring i uka utenfor den vanlige barnehagegruppa. Det har ikke blitt forsket på utkommet av denne opplæringa, men det

er klart at denne modellen ikke vil være tilstrekkelig til å oppnå gode samiskkunnskaper i løpet av barnehagetida. Den kan ha en viss positiv effekt på lang sikt, som forberedelse til samiskopplæring i grunnskolen. En annen mulighet er at den kan støtte opp om samiskopplæringa i hjemmet dersom barnet har folk rundt seg som snakker samisk. Modell 5, majoritetsspråkmodellen, innebærer at det ikke er noen opplæring i samisk. Som det ble sagt i begynnelsen av dette underkapittelet, er det all grunn til å tro at svært mange samiske barn går i en modell 5 barnehage der de ikke har noen opplæring i samisk. Dermed lærer de ikke mer samisk i barnehagen enn det minimumet som alle barn i landet skal lære.

5 De samiske barnehagenes historikk

Opprettelsen av de samiske barnehagene kan sees som et svar på ei samfunnsutvikling fra 1960-tallet der det blir vanligere at begge foreldre jobber utenfor hjemmet. Når samiske foreldre trengte barnehageplass til barna, ville de ha et tilbud på samisk, med samisk innhold. Et ønske om at barnehagen også skal bidra til revitalisering av språket, kommer også tidlig inn. Det er helt klart til stede ved opprettelse av barnehager i Skånland og Tysfjord i Norge rundt 1990, og det altoverveiende argumentet når skoltesamene og enaresamene oppretter språkreir på 1990-tallet. Det blir uten tvil stadig viktigere også utover 2000-tallet. Det ser man bl.a. av den kraftig økning i antallet språkreir i Finland på 2000-tallet. (Storjord 2004; 2008: 39–46; Laiti 2018; Nuorgam 2019; Sametinget i Finland 2020; 2022; Sametinget i Sverige 2013; 2021; Eriksen 2019/2020; Kuhmunen 2019/2020).

5.1 Samiske barnehager på norsk side

Den aller første samiske barnehagen ble opprettet i 1969 i Kautokeino. Driftsspråket var imidlertid norsk de første årene (Storjord 2008: 41). Deretter foregikk det ei langsom etablering av samiske barnehager andre steder. Først ble det etablert samiske barnehager på 1970-tallet i Máze, Mironjávri, Karasjok, Tana og Porsanger, deretter i Nesseby, Tromsø, Kåfjord, Vadsø, Skånland og Oslo på 1980-tallet. I 1989 ble den første lulesamiske barnehagen etablert i Tysfjord (nå Hamarøy kommune). På 1990-tallet ble gjort flere forsøk på å etablere sørsamiske barnehagetilbud uten at det lyktes å få dem til å bli varige. På 1990-tallet ble det etablert flere barnehager i kommuner som allerede hadde slike tilbud fordi behovet økte (bl.a. Karasjok, Tana, Porsanger og Tromsø) og det ble etablert nye samiske barnehager (f.eks. i Lavangen og Musken i Tysfjord). To trekk er spesielle for 2000-tallet. Det ene er at de etablerte samiske barnehagene i byene har stor pågang og utvider tilbudet. Det andre trekket er at det blir etablert sørsamiske barnehager som faste tilbud: På Snåsa i 2003, senere i Røyrvik, på Røros og i Trondheim. (Storjord 2004; 2006; 2008: 39–46; Eriksen 2019/2020).

I 2020 fantes 52 barnehager med tilbud om samisk språk i en eller annen form, og 813 barn som tok del i disse tilbudene. Det store flertallet av barna, 711, gikk i en av de 32 samiske barnehagene eller samiske avdelingene i norskspråklige barnehager, mens 102 barn fikk tilbud om samisk språk i totalt 20 norske barnehager. De fleste samiske barnehagene lå i 2020 i det nordsamiske området. To av barnehagene var i lulesamisk område og fire av barnehagene var i sørsamisk område. Det var samiske barnehager i fem byer: Alta, Tromsø, Bodø, Trondheim og Oslo. (SSB 2021; Sametinget i Norge 2021).

5.2 Samiske barnehager på svensk side

Den eldste samiske barnehagen i Sverige er Skierri i Gällivare. Den ble opprettet i 1986. En samisk fulltidsbarnehage ble etablert i Kiruna i 1987 etter å ha vært et deltidsprosjekt i tre år. Detter ble det opprettet samiske barnehager i Jokkmokk i 1992, i Karesuando i 2003 og i Tärnaby i 2010. Antallet barn i disse barnehagene var 120 i 2021. Alle disse barnehagene er underlagt og driftes av den Svenske sameskolstyrelsen. De nordligste barnehagene har nordsamisk som driftsspråk, i Jokkmokk og Gällivare er drifta både på nord- og lulesamisk, mens tilbudet i Tärnaby er på sørsamisk. Bruken av samisk språk i Sameskolestyrelsens samiske barnehager varierer. Det fins avdelinger ved de samiske førskolene der majoriteten av barna ikke har samisk som førstespråk. Det innebærer at kommunikasjonsspråket tenderer til å bli svensk. For personalet innebærer det vanskeligheter med å stagge og snu bruken av svensk i barnehagen. Ved barnehager der det er vanskelig å finne pedagoger med samisk språkkunnskap, blir barna

ikke naturlig samiskspråklige i løpet av tiden i barnehagen. (Stockholms län/Sametinget i Sverige 2013; 21–22; Sametinget i Sverige 2013; 2021: 19–23; Kuhmunen 2019/2020).

I 2010 ble tretten nye kommuner innlemmet i forvaltningsområdet for samisk språk i Sverige. Ingen av dem hadde samisk barnehagetilbud før innlemmelsen, men nå fikk de ansvar for å tilby barnehagetilbud på samisk[6]. De fleste av disse kommunene har nå et samiskspråklig barnehagetilbud, men tilbudet er begrenset fra en klokketime i uka til tre dager i uka. Det er ikke tilgjengelige tall for antallet barn som deltar i disse tilbudene. En henvendelse fra forfatteren av denne artikkelen til kommunene resulterte i så få svar at ingen konklusjoner kan trekkes. (Stockholms län/Sametinget 2013: 40; Sametinget i Sverige 2021: 19–23).

5.3 Samiske barnehager på finsk side

På finsk side prøvde kommunene fram til 1990-tallet, og i noen tilfeller ut på 2000-tallet, å tilby et tospråklig finsk/samisk tilbud der noen, gjerne bare én ansatt, skulle snakke samisk med de samiskspråklige ungene. Dette førte til at finsk dominerte i barnehagene. Det skjedde imidlertid ei endring utover på 1990-tallet der kommunene opprettet samiskspråklige barnehager i *Samernas hembygdsområde* og seinere utenfor dette området. Den første samiskspråklige barnehagen på finsk side ble opprettet i Utsjoki i 1991. Etter dette har det blitt opprettet samiskspråklige barnehager i Ivalo i 1995, i Inari i 1996, i Hetta i 1998 og Karigasniemi i 1999. Utover 2000-tallet ble det opprettet samiskspråklige barnehager i Oulu i 2010, Rovaniemi og Helsingfors i 2013 og i Kittilä i 2018. I 2020 var det 96 barn i ni samiskspråklige barnehager som alle hadde nordsamisk som driftsspråk. (Raudasjoki 2016: Sametinget i Finland 2020).

I tillegg til barna i samiskspråklige barnehager, var det 80 barn i 12 samiske språkreir. Etableringa av disse språkreirene har sin spede begynnelse på 1990-tallet da språkreirmetodikken ble tatt i bruk for skoltesamisk og enaresamisk. Fram til 2010 var det vanskelig å finansiere språkreirene og kun ett av dem, et enaresamiske språkreir i Enare, har vært kontinuerlig i drift siden starten i 1997. Det er nå skoltesamiske språkreir i drift i Sevettijärvi og Ivalo, to enaresamiske språkreir i Enare og ett i Ivalo. Det første nordsamiske språkreiret startet i Vuotso i 2007. Deretter ble språkreir opprettet i Utsjoki og Karigasniemi i 2012, i Helsingfors i 2013 og i Oulu, Rovaniemi og Sodankylä i 2015. I 2018 ble språkreir etablert i Nuorgam i Utsjoki kommune og i Hetta i Enontekio. (Raudasjoki 2016; Nuorgam 2019; Sametinget i Finland 2020). Piia Nuorgam (2019: 8), som utredet behovet for samiske språkreir for Sametinget i Finland, konkluderte med at behovet var nesten dekket i *Samernas hembygdsområde,* mens 92,5 prosent av de samiske barna lenger sør manglet et samisk barnehagetilbud.

5.4 Samisk barnehage på russisk side

Det fins et samiskspråklig tilbud for barnehagebarn i Russland. Dette har eksistert siden 1994 på kildinsamisk i Luujäu′rr/Lovozero. Her ble ei samisk barnehageavdeling opprettet ved en barnehage etter krav fra foreldrene, og 15 barn tok del i samiskopplæringa. Samisk språk har aldri vært hovedspråk på avdelinga, men samisk ble brukt to timer om dagen i perioden 1994–2008. Etter dette har bruken av samisk minket. I perioden 2018–2020 var det opplæring i samisk 20 minutter fire ganger i uka. Tilbudet stoppet i 2020 pga. koronapandemien. Det har ikke vært mulig å få informasjon om antallet barn som har mottatt opplæring i samisk i barnehagen de senere år. (Jushkov 2020; Vasilijeva 2020).

6 Diskusjon: Barnehager, språk og undervisningsmodeller

Det er absolutt et positivt trekk at mer enn tusen barn har plass i samiske barnehager der samisk er hovedspråk, og at mer enn hundre barn får tilbud om opplæring i samisk språk i andre barnehager. Det er også et positivt trekk at det i tillegg til nordsamisk, også gis tilbud om opplæring i kildin-, skolte-, enare-, nord-, lule- og sørsamisk i samiske barnehager.

Sett fra en språksosiologisk synsvinkel er det viktig å kunne overvåke språksituasjonen for et minoritetsspråk for å kunne utvikle en språkpolitikk som treffer og forbedrer språkenes vitalitet. En av

[6] Lag (2009: 724) om nationella minoriteter och minoritetsspråk § 17. https://lagen.nu/2009:724. Se også Skollag (2010: 800) kap 8. § 12a. https://lagen.nu/2010:800#K8

nestorene innen forskningsfeltene revitalisering av minoritetsspråk og vending av språkskiftet, Joshua Fisman, advarte tidlig mot å miste fokuset på barnas språktilegnelse fordi man blir så blendet av framskritt for minoritetsspråket innen media, administrasjon og andre offentlige arenaer. Han rådet minoritetene til også å ha et konstant søkelys på overføring av språket til nye generasjoner. (Fishman 1991: 107). Innsamling og analyse av empirisk, kvantitativt materiale om språkbruk og språkopplæring i barnehagene kan derfor være et svært et svært viktig bidrag til å holde søkelyset på vitaliteten til de samiske språkene. Innsamlinga av empirisk materialet til denne artikkelen har avdekket både positive sider og en del svakheter ved mulighetene for å finne relevant informasjon. Det positive er at Sametingene i Norge og Finland har god oversikt over antall barn i samiske barnehager fordi de forvalter støtteordninger til disse. Lignende tall fra Sverige viste det seg å være vanskelig å samle inn, fordi det ikke er noe offentlig organ som har den totale oversikten.

Sametinget i Norge forvalter også tilskudd til samisk språkopplæring for barn i norske barnehager. Her er det tilgjengelige tall for antall barnehager som gir tilbudet, og antall barn som mottar tilbudet. Men det ikke mulig å finne noen opplysninger om de språkopplæringsmetodene som brukes, eller om resultatet av denne opplæringa. Også i Sverige har barn rett til ei slik opplæring, men tallmateriale mangler. Ifølge Baker (2001) er ei slik opplæring i utgangspunktet en svak språkopplæringsmodell der resultatet sannsynligvis er enspråklighet eller begrenset tospråklighet. Dette kan likevel betraktes som ei akseptabel løsning når det ikke er mulig å tilby sterkere språkopplæringsmodeller i samisk språk.

Fordi Sametinget i Finland i sine støtteordninger skiller mellom samiskspråklige barnehager og samiske språkreir, har de også kunne levere god statistikk for dette feltet. Sametinget i Norge opererer ikke med et slikt skille i sine støtteordninger, og dermed kan vi ikke si noe sikkert om språkopplæringsmodellene i de samiske barnehagene på grunnlag av den tilgjengelige informasjon. Det vi vet utfra betingelsene for å bli støtteberettiget som samisk barnehage eller samisk barnehageavdeling i Norge, er at disse samiske barnehagene og barnehageavdelingene skal bruke en av Bakers sterke språkopplæringsmodeller. Dette er positivt for språkenes vitalitet. Men det hadde vært en fordel å se hvordan antallet barn fordeler seg mellom førstespråksmodellen, språkreirsmodellen og eventuelt andre språkopplæringsmodeller, for å kunne vurdere om dette forholdet endrer seg over tid.

Utfra beregninger av den samiske befolkningas størrelse må en kunne slå fast at de samiske barna som faktisk har et samiskspråklig tilbud i barnehagen, bare utgjør en del av de samiske barna. De fleste samiske barna vil en finne i norske, svenske, finske og russiske barnehager der de ikke har tilbud om samisk språkopplæring. Dette klassifiserer Baker som en svak opplæringsmodell der det sannsynlige utkommet er enspråklighet i majoritetsspråket. Som vist i kapittel 5, er det kun i Finland at Sametinget har gjort beregninger av hvor stor del av de samiske barna som har plass i en samisk barnehage. For de andre landene mangler vi empiri, og det er en svakhet for overvåkinga av de samiske språkenes vitalitet.

Forskninga på samiske barnehager gir en del viktig informasjon om språkbruk og språkopplæringsmodeller i barnehagene. En del av forskningslitteraturen viser at barnehagene bruker førstespråksmodellen og språkreirsmodellen som ifølge Baker er sterke språkopplæringsmodellene med tospråklighet som resultat (Storjord 2004; 2008; Todal 2007; Øzerk 2006; Braut 2010; Laiti 2018; Äärälä-Vihriälä 2016; Pasanen 2010; 2015; 2018; Rasmussen 2013a; Laihi 2017; Kleemann 2015; 2021). Imidlertid kommer det også fram at noen barnehager bruker en blandingsmodell der samiskspråklige og ikke-samiskspråklige barn går i samme barnehage. Dette er ifølge Baker en sterk språkopplæringsmodell, med tospråklighet som utkomme, når den fungerer. Men i eksemplene fra forskning på samiske barnehager var blandingsmodellen tatt i bruk uten ei planmessig og differensiert opplæring i samisk for de barna som ikke kunne språket i utgangspunktet. Resultatet var at noen barn ikke lærte å snakke samisk, og barna snakket mest norsk med hverandre i barnehagen. (Storjord 2008). Det må legges til at kun noen få barnehager er undersøkt i et relativt kort tidsrom. Derfor kan vi ikke generalisere på grunnlag av de forskningsresultatene.

Forskningsresultatene fra finsk side viser at en bevisst har bygd opp et todelt system der samiskspråklige barnehager er for samiskspråklige barn, mens samiske språkreir er et tilbud til ikke-samiskspråklige barn. Heller ikke her har vi mange forskningsresultater å vise til. Men det som finnes om enaresamiske språkreir, viser svært gode resultat når det gjelder barnas språkinnlæring og språkbruk. (Pasanen 2010; 2015; 2018). Slik forskning er for øvrig en mangelvare. Vi kan ikke vise til forskning som viser resultatet av de samiske barnehagene eller samiskopplæring i andre barnehager når det gjelder barnas samiske språk-

kunnskaper og språkbruk. Derfor er det vanskelig å si hvor godt barnehagene lykkes med sin språkopplæring. Forskning på dette burde være et satsingsområde. Noen av pedagogikkforskerne har forsket på hvordan språkopplæringa foregår i språkreir (Äärälä-Vihriälä 2016; Laihi 2017; Kleemann 2021). Dette er forskning det absolutt er mer behov for, ettersom eksempler på god praksis kan være til stor hjelp for andre som arbeider med barn i språkreir, og det kan danne grunnlag for forskningsbasert undervisning i barnehagelærerutdanningene.

7 Konklusjon

Denne artikkelen behandler det samiske barnehagefeltet der det ble undersøke hvilke språkopplæringsmodeller som brukes i barnehagene for å utvikle barns samiske språk, og det ble stilt spørsmål om det tilgjengelige tallmaterialet kan danne grunnlag for å vurdere samiske språks vitalitet. Artikkelen viser at det er problematisk å få en oversikt over hele feltet som geografisk spenner over fire land. Dette kommer av at det er ulik registreringspraksis eller mangel på registreringspraksis i landene. Dette gjør det igjen vanskelig å bruke tall fra de samiske barnehagene i overvåkinga av de samiske språkenes vitalitet. Dette problemet kan løses hvis Sametingene og de relevante, sentrale myndighetene i landene vil samarbeide om det.

I artikkelen blir de språkopplæringsmodellene som brukes i samiske barnehager og for samiske barn i andre barnehager, klassifisert på grunnlag av faglitteratur om språkopplæringsmodeller og forskning på samiske barnehager. Dette er ment som et bidrag til en debatt om samiske språkopplæringsmodeller i barnehagen. Håpet er at dette kan være med å styrke de samiske barnehagene og hjelpe flere barn til å bli samiskspråklige.

Artikkelen viser også at forskere både med utgangspunkt i språksosiologi og barnehagepedagogikk har bidratt med ny kunnskap om den samiske barnehagen. Den er likevel et relativt lite utforsket felt der vi trenger bidrag fra begge gruppene forskere for å øke antallet samiskspråklige og sikre ei framtid for samiske språk.

Referanser

Baker, Colin. 2001. *Foundation of Bilingual Education and Bilingualism*. Tredje utgave. Multilingual Matters, Clevdon.

Braut, Kristine Tjåland. 2010. *To Speak or Not to Speak. Because they tell me to speak Sámi at daycare. Indigenous language revitalization through preschool children learning a second language in a language nest.* Masteroppgave, Universitetet i Tromsø. Tilgjengelig på https://hdl.handle.net/10037/2965.

Eriksen, Anne Ingebjørg Svineng. 2019/2020. Lærer på samisk barnehageutdanning ved Samisk høgskole. Personlig kommunikasjon gjennom skoleåret 2019/2020.

Fishman, Joshua A. 1991. *Reversing Language Shift. Theoretical and Empirical Foundation of Assistance to Thretened Languages.* Multilingual Matters 76, Clevdon. https://doi.org/10.2307/330061.

Guttorm, Juha. 1986. *Alle kouluikäisten saamelaisten kasvuolosuhteet ja niiden kehittämismahdollisuudet Utsjoen kunnassa.* [Oppvekstvilkår for samiske barn under skolealder og deres utviklingsmuligheter i Utsjoki kommune.] Valtion painatuskeskus, Helsingfors.

Hinton, Leanne og Ken Hale. 2001. *The Green Book of Language Revitalization in practice*. Academic Press, New York. https://doi.org/10.1163/9789004261723.

Huss, Leena. 1999. *Reversing Language Shift in the Far North. Linguistic Revitalization in Northern Scandinavia and Finland.* Studia Uralica Upsaliensia 31, Uppsala.

Hyltenstam, Kenneth og Christopher Stroud. 1991. *Språkbyte och språkbevarande. Om samiskan och andra minoritetsspråk.* Studentlitteratur, Lund.

Hyltenstam, Kenneth, Christopher Stroud og Mikael Svonni.1999. Språkbyte, språkbevarande och revitalisering. Samiskans ställning i svenska Sápmi. I *Sveriges sju inhemska språk – ett minoritetsspråksperspektiv*, redigert av Kenneth Hyltenstam, s. 41–97. Studentlitteratur, Lund.

Ijäs, Johanna Johansen. 2011. *Davvisámegiela finihtta vearbahámiid sojahanvuogádaga oččodeapmi vuollel golmmajahkásaš máná gielas*. [Tilegnelse av nordsamiskspråklige finitte verbformers bøyningsmønstre i språket til et barn under tre år]. Ph.D. avhandling, Davvi Girji, Kárášjohka. Tilgjengelig på http://hdl.handle.net/11250/177041.

Jansson, Annika. 2005. *Sami Language at Home and at School. A Fieldwork Perspective*. Studia Uralica Upsaliensia 36, Uppsala.

Johansen, Åse Mette. 2009. *"Velkommen te' våres Norge." En kvalitativ studie av språkbytte og språkbevaring i Manndalen i Gáivuotna/Kåfjord*. Novus Forlag, Oslo.

Jushkov, Jevgenij. 2020. *Sámi beaiveruoktu Ruoššas.* [Samiske barnehager i Russland.] Et upublisert notat om samiske barnehager i Russland samt om lovverket som regulerer samiske barnehager i Russland. Datert 4. august 2020.

King, Kendall A. 2009. Language loss and revitalization: Ten things we know. I *Kvener og skogfinner i fortid og nåtid*, redigert av Anna-Riitta Lindgren, s. 9–23. Speculum Boreale, Tromsø.

Kleemann, Carola. 2015. *Lek på to språk. En studie av kodeveksling og språkalternering i tospråklig rollelek på nordsamisk og norsk i en samisk barnehage*. Ph.D. avhandling, UiT Norges arktiske universitet, Tromsø. Tilgjengelig på https://hdl.handle.net/10037/8153.

Kleemann, Carola. 2021. Pedagogical Translanguaging to Create Sustainable Minority Language Practices in Kindergarten. *Sustainability* 13(7): 3613. https://doi.org/10.3390/su13073613.

Kuhmunen, Gudrun. 2019/2020. Lærer på samisk barnehageutdanning ved Samisk høgskole. Personlig kommunikasjon gjennom skoleåret 2019/2020.

Laihi, Tiina-Maaria. 2017. *Skolt Sámi Language and Cultural Revitalization : A case study of a Skolt Sámi language nest.* Pro gradu-oppgave, Helsingfors universitet, Helsingfors. Tilgjengelig på http://urn.fi/URN:NBN:fi:hulib-201706194962.

Laiti, Marikaisa. 2018. *Saamelaisen varhaiskasvatuksen toteutus Suomessa.* [Implementering av samisk førskoleopplæring i Finland.] Ph.D. avhandling, Universitetet i Lappland, Rovaniemi. Tilgjengelig på https://urn.fi/URN:ISBN:978-952-337-098-2.

Laiti, Marikaisa. 2019. History of Early Childhood Education in the Sámi Language in Finland. I *Sámi Educational History in a Comparative International Perspective*, redigert av Otso Kortekangas, Pigga Keskitalo, Jukka Nyyssönen, Andrej Kotljarchuk, Merja Paksuniemi og David Sjögren, s. 187–206. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-24112-4_11.

Nuorgam, Piia. 2019. *Saamelaiskäräjien mahdollisuudet hallinnoida ja järjestää saamelaisten kulttuuri- ja kielipesätoimintaa*. Esiselvityshanke. [Sametingets muligheter til å forvalte og organisere de samiske kultur- og språkreirsvirksomhetene. Forstudieprosjekt.] Sametinget i Finland, Enare.

Pasanen, Annika. 2010. Will language nests change the direction of language shifts? On the language nests of Inari Saamis and Karelians. I *Planning a new standard language: Finnic minority languages meet the new millenium,* redigert av Harri Sulkala og Helena Mantila. Finska litteratursälskapet, Helsingfors.

Pasanen, Annika 2015. *Kuávsui já peeivičuovâ. 'Sarastus ja päivänvalo': Inarinsaamen kielen revitalisaatio* [Kuávsui já peeivičuovâ. 'Daggry og dagslys': Revitalisering av det enaresamiske språket]. Ph.D. avhandling, Helsingfors universitet, Helsingfors. Tilgjengelig på https://docplayer.fi/35092451-Kuavsui-ja-peeivicuova-sarastus-ja-paivanvalo-inarinsaamen-kielen-revitalisaatio.html.

Pasanen, Annika. 2018. "This Work is Not for Pessimists": Revitalization of Inari Sami Language". I *Routledge Handbook of Language Revitalization*, redigert av Leanne Hinton, Leena Huss og Gerald Roche, s. 364–372. Routledge, Abingdon. https://doi.org/10.4324/9781315561271-46.

Rasmussen, Torkel. 2005. *Jávohuvvá ja ealáska: Davvisámegielagiid demografiija ja buolvvaidgaskasaš sirdáseapmi Norggas ja Suomas*. [Stilner og livner til. Nordsamisk taleres demografi og intergenerasjonell språkoverføring i Norge og Finland.] Hovedfagsoppgave, Universitetet i Tromsø, Tromsø. Tilgjengelig på https://hdl.handle.net/10037/5064.

Rasmussen, Torkel. 2013a. *"Go ealáska, de lea váttis dápmat" : Davvisámegiela etnolingvisttalaš ceavzinnávccaid guorahallan guovtti gránnjágielddas Deanus ja Ohcejogas 2000-logu álggus.* ["Når den livner til, er den vanskelig å temme" : En undersøkelse av den etnolingvistiske vitaliteten for nordsamisk i to nabokommuner Tana og Utsjoki på begynnelsen av 2000-tallet.] Ph.D. avhandling, UiT Norges arktiske universitet, Tromsø. Tilgjengelig på https://hdl.handle.net/10037/5593.

Rasmussen, Torkel. 2013b. Sametingets støtteordninger til samiske språk. I *Samiske tall forteller 6. Kommentert samisk statistikk*, s, 49–72. Samisk høgskole, Kautokeino. Tilgjengelig på https://samilogutmuitalit.no/nb/2013/sametingets-midler-til-samiske-sprak.

Raudasjoki, Tea. 2016. *"Kultuvra stivre min doaimmaid". Sámi mánáidgárdeoahpaheddjiid vásáhusat sámemánáidgárddi bargobeaivvis.* [Kulturen styrer våre aktiviteter. Samisk barnehagelæreres erfaringer fra hverdagen i samiske barnehager]. Bacheloroppgave, Samisk høgskole, Kautokeino.

Sametinget i Finland. 2020. *Mánáidlohku Sápmelaš árrabajásgeassimis 2012–2020.* [Barnetallet i samiske barnehager 2012–2020]. En oversikt oversendt fra Sametinget i Finlands administrasjon. Sametinget i Finland, Enare.

Sametinget i Finland. 2022. *Sámiid kultur- ja giellabeassedoaibma Suomas* [Samisk kultur og språkreiraktivitet i Finland]. Sametinget i Finland, Enare. Tilgjengelig på https://www.samediggi.fi/doaibma/samegiella/giellabeassi/samiid-kultur-ja-giellabeassedoaibma-suomas/?lang=dav.

Sametinget i Norge. 2021. *Oversikt over hvilke barnehager som har fått innvilget støtte i 2020.* Sametinget i Norge, Karasjok. Tilgjengelig på: https://sametinget.no/_f/p1/ib4be59c9-1ad4-47d0-a09a-16478ca8f553/oversikt-hvilke-barnehager-som-har-fatt-innvilget-stotte-til-samisk-barnehage-og-samisk-avdeling-i-norske-barnehager-i-2020.pdf%20Lest%2011.12.%202021. Lest 11.12. 2021.

Sametinget i Norge. 2022. *Regelverk samiske barnehager og barnehager med samisk avdeling - søkerbaserte tilskudd 2022.* Sametinget, Karasjok. Tilgjengelig på https://sametinget.no/stipend-og-tilskudd/oversikt-over-tilskuddsordninger/barnehage/tilskudd-til-samiske-barnehager-og-samisk-avdeling-i-norsk-barnehage.

Sametinget i Sverige. 2011. *Samisk utbildningspolitik – en introduktion. En rapport fra Sametingets utbildningskommitté*. Sametinget i Sverige, Kiruna. Tilgjengelig på chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fwww.sametinget.se%2F31895&pdffilename=Rapport%20%20Samisk%20utbildningspolitik_110822.pdf.

Sametinget i Sverige. 2013. *Lägesrapport: De samiska språken i Sverige 2013.* Sametinget i Sverige, Kiruna. Tilgjengelig på http://www.sametinget.se/70485.

Sametinget i Sverige. 2021. *Lägesrapport: De samiska språken i Sverige 2021*. Sametinget i Sverige, Kiruna. Tilgjengelig på https://www.sametinget.se/lagesrapporter_sprak.

Samisk høgskole. 2022. *Sámi mánáidgárdepedagogihkka ođđa áiggis.* [Samisk barnehagepedagogikk i ei ny tid]. Samisk høgskole, Kautokeino. Tilgjengelig på https://samas.no/se/a/dutkan/proseavttat/sami-manaidgardepedagogihkka-odda-aiggis.

Sammallahti, Pekka. 1998. *The Saami Languages. An Introduction*. Davvi Girji, Karasjok.

SSB. 2021. *Samiske barnehager, barnehager med samisk språkopplæring, barn i samiske barnehager og barn med samisk språkopplæring 2005–2020*. Statistisk sentralbyrå, Oslo. Tilgjengelig på https://www.ssb.no/statbank/table/06777/.

Stockholms län/Sametinget i Sverige. 2013. *Nationella minoriteter : Rapport om tillämpningen av lagen om nationella minoriteter och minoritetsspråk år 2013*. Stochholms län og Sametinget i Sverige. Stockholm/Kiruna. Tilgjengelig på http://sametinget.se/71367?file_id=1.

Storjord, Marianne Helene. 2004. *«Fra nødhjelp til folkehjelp.» Opprettelse og utbygging av samiske barnehager i Norge i 1969–99.* Masteroppgave, Universitetet i Tromsø, Tromsø.

Storjord, Marianne Helene. 2006. Samiske barnehagers historie. I *Samisk skolehistorie* 3*,* redigert av Svein Lund, Elfrid Boine, Siri Broch Johansen og Siv Rasmussen. Davvi Girji, Karasjok. Tilgjengelig på http://skuvla.info/skolehist/storjord-n.htm.

Storjord, Marianne Helene. 2008. *Barnehagebarns liv i en samisk kontekst – En arena for kulturell meningsskaping.* Ph.D. avhandling, Universitetet i Tromsø, Tromsø. Tilgjengelig på https://hdl.handle.net/10037/3571.

Todal, Jon. 2002. *"... jos fal gáhttet gollegielat". Vitalisering av samisk språk i Noreg på 1990-talet.* Ph.D. avhandling, Universitetet i Tromsø, Tromsø. Tilgjengelig på https://hdl.handle.net/10037/12050.

Todal, Jon. 2007. *Samisk språk i Svahken Sijte. Sørsamisk vitalisering gjennom barnehage og skule.* Dieđut nr. 1/2007. Samisk høgskole, Kautokeino.

UNESCO. 2003. *Language Vitality and Endangerment*. UNESCO Ad hoc Group on Endangered Languages, Paris.

Vasilieva, Alla. 2020. Tidligere samisklærer i barnehagen *Solnyško (Sola)* i *Luujäu'rr*/Lovozero. Intervjuet av Jevgenij Jushkov sommeren 2020. Se Jushkov 2020.

Äärälä-Vihriälä, Rauni. 2016. *"Dat ii leat dušše dat giella" – "Se ei ole vain se kieli" : tapaustutkimus saamenkielisestä kielipesästä saamelaisessa varhaiskasvatuksessa.* [«Dat ii leat dušše dat giella» - «Det er ikke bare det språket»: en casestudie av et samiske språkreir i samisk småbarnspedagogikk.] Ph.D. avhandling, Universitetet i Lappland, Rovaniemi. Tilgjengelig på https://urn.fi/URN:ISBN:978-952-484-929-6.

Cyclic feeding interactions between finite-state mal-rules: an algorithm for the optimal grouping and ordering of mal-rules

Robert Reynolds[†‡], Laura Janda[†], Tore Nesset[†]
[†] *UiT–Arctic University of Norway*, *Tromsø, Norway*
[‡] *Brigham Young University*, *Provo, UT, USA*

Abstract

Intelligent Language Tutoring Systems typically attempt to automatically diagnose learner errors in order to provide individualized feedback. One common approach is the use of mal-rules to extend normative grammars by licensing specific types of learner errors. In finite-state morphologies, mal-rules can be implemented as two-level rules or replace rules. However, unlike the phonological rules of natural languages, mal-rules do not necessarily behave as a coherent system, especially with respect to feeding interactions. Using examples from learner errors attested in the RULEC corpus of Russian learner texts, we illustrate the problem of cyclic feeding interactions that can occur between mal-rules. We then describe a formal algorithm for identifying an optimal ordering for mal-rules to be applied to a transducer.

Keywords: Learner errors, mal-rules, Russian, rule ordering, finite-state transducer

1. Introduction

Intelligent Language Tutoring Systems (ILTS) automatically analyze language produced by a learner in order to provide individualized feedback. Examples of ILTS include E-Tutor (Heift 2010), Robo-Sensei (Nagata 2009), TAGARELA (Amaral and Meurers 2011), the i-tutor (Choi 2016), the FeedBook (Meurers et al. 2019). Unlike a spell-checker, which merely suggests correct forms, an ILTS can provide remedial explanation and examples to help the learner understand *why* a given form is incorrect, or *how* to correct the error. This article is connected to recent work on a new ILTS: Russian Mentor for Orthographic Rules (RuMOR) (Reynolds et al. 2022).[1]

In order to provide this information, an ILTS must "abstract away from the specific string entered by the learner to more general classes of properties by automatically analyzing the learner input using NLP algorithms and resources" (Meurers 2020). This article is concerned with one particular approach to learner language analysis: mal-rules. Mal-rules are purposefully mal-formed rules that produce the same kinds of errors that learners make. By applying such rules to existing analyzers or parsers, we license learner errors, which can then be recognized by the system. In this way, we are able to analyze structures that are absent from normative Natural Language Processing (NLP) systems.

Mal-rules can interact with other rules in ways that violate the coherence of the rule system as a whole. This can pose technical issues for systems that integrate a large number of mal-rules. A simple demonstration of this problem is misspellings caused by learners' failure to distinguish two sounds in the target language. For example, many learners of Russian do not distinguish between the letters ш and щ (*š* and *šč* in transcription) or their respective sounds, [ʂ] and [ɕː]. This confusion would lead to the following errors: *xoroščo* (c.f., *xorošo* 'good') or *piša* (c.f., *pišča* 'food'). Modeling these kinds of errors might require two rules, as shown in (1).

(1) a. šč → š (and add the tag +shch2sh)
b. š → šč (and add the tag +sh2shch)

Applying these rules sequentially to the word *pišča* would result in the analyses shown in (2).

(2) a. *pišča* : *pišča*
b. *pišča*+shch2sh : *piša*
c. *pišča*+shch2sh+sh2shch : *pišča*

As can be seen in (2), applying rule (1a) to (2a) results in (2b), which is the expected error, with the appropriate tag. However, because the first rule feeds into the second, when rule (1b) is subsequently applied to the lexicon, it

[1] https://icall.byu.edu/rumor

 Nordlyd 46.1: 219—229, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6306

converts (2b) back into the original, correct wordform (note that (2b) and (2c) have identical surface forms). The result is an analyzer that would return two analyses for the surface form *pišča*, one with its correct normative reading, and one with two contradictory error tags. Reversing the order of the rules would fix the problem for *pišča*, but then the word *xorošo* would have an analogous problem. We refer to this set of circumstances as a cyclic feeding interaction. Although two-rule cycles, such as this example, are the most common, cyclic feeding interactions can involve any number of rules. Furthermore, a given rule can be part of any number of cycles.

Our use of the term *cycle* is not connected to uses in Cyclic Phonology and related generative frameworks. Their use of the term *cyclic* stems from the work of Chomsky et al. (1956) in their analysis of English word stress. In research inspired by their work, cycles are iterative application of the same rule(s) at increasingly larger morphosyntactic units. In contrast, our use of the term *cycle* is taken from Graph Theory, as discussed in Section 2.2. We represent rules in a graph, where each rule is a node in the network, and feeding interactions between rules are represented as directed edges (arrows) between nodes. In this representation, a cycle is a relation between rules where feeding interactions ultimately lead back to the original form. As discussed in Section 2.1, some researchers name this a "mutual feeding" relation, but we avoid this term because it implies that only two rules are involved in the interaction.

1.1. Acyclic feeding interactions

We have seen that cyclic feeding relations can be problematic, but *acyclic* feeding interactions are also consequential for rule interactions. The toy rules in (3) are an example of rules in a feeding order. Because they are in a feeding order, the analyzer would recognize wordforms with both errors on the same word (e.g. cat+a2o+o2u : cut). However, if the order of the rules is reversed (a counterfeeding order), the analyzer would fail to recognize this interaction of errors. In order to model mal-rule interactions, it is important to optimize the order of all the rules to maximize the number of rules in feeding orders (and minimize the number of rules in counter-feeding orders). Given that a set of n rules can be arranged in $n!$ different permutations, an algorithm that can approach this task automatically would be a boon to rule authors.[2]

(3) a. a → o (and add the tag +a2o)

b. o → u (and add the tag +o2u)

1.2. Article structure

In this article, we present a formal algorithm, implemented as an open-source python script,[3] to block cyclic feeding interactions. In addition, our algorithm assesses the consequences of ordering *acyclic* feeding interactions, and suggests an ordering of errors to optimize the feeding interactions of such rules.

In Section 2, we discuss related research in three fields: generative phonology (§2.1), mal-rules (§2.3), and Graph Theory (§2.2). In Section 3, we describe the algorithm. In Section 4, we apply the algorithm to a real-world set of mal-rules modeling Russian learner orthographic and morphological errors. Finally, in Section 5, we summarize our contribution and outline a few paths for future work.

2. Related work

2.1. Generative approaches to feeding and bleeding orders

The terms *feeding order* and *bleeding order* were first introduced by Kiparsky (1968). A feeding order is an ordering of two rules such that the first rule generates new contexts in which the second order applies. Stated negatively, for some words the second rule would not apply if the first rule had not created the right context. When two rules could have a feeding order, but they are not in the right order for a feeding interaction to occur, they are said to be in a *counterfeeding*

[2] It is worth noting that rule interactions of this kind can rapidly explode the size of a transducer, which could strain computational resources. In this case, one might wish to *avoid* modeling mal-rule interactions in order to keep the transducer smaller. This can be achieved by reversing the optimal feeding order to produce the optimal counter-feeding order.

[3] https://github.com/reynoldsnlp/xfst_malrule_ordering

order. Whereas a feeding order is a *timely* application of a feeding interaction, a counterfeeding order can be said to be a *tardy* application of a feeding interaction.

A bleeding order is an ordering of two rules such that the first rule destroys contexts in which the second rule applies. In other words, for some words the second rule would apply if not for the first rule. Just as with feeding interactions, if two rules could have a bleeding order, but they are in the wrong order for a bleeding interaction to occur, they are in a *counterbleeding* order. A summary of these interactions is given in Table 1. In this article, we focus only on *feeding* interactions.

		Chronology	
		timely	tardy
Interference	excitatory	feeding	counterfeeding
	inhibitory	bleeding	counterbleeding

Table 1: Types of rule interaction

Generative linguists have wrestled with the problem of *cyclic* feeding orders, albeit for different reasons than the present article. For their theories, cyclic feeding interactions (also known as "mutual" feeding interactions) pose problems for explanatory power, parsimony (i.e. Occam's Razor), and learnability. For example, one mechanism that can be used to solve the problem of cyclic/mutual feeding is disjunctive ordering (Chomsky and Halle 1968), which uses braces/brackets/parentheses to define a schema from which only one rule can be applied. Similarly, the Elsewhere Condition (Kiparsky 1973) was put forth as an alternative to disjunctive ordering.

Pullum (1976) expresses suspicion of derivations of the type A → B → A, which he calls "Duke-of-York derivations."[4] He argues that in most cases, Duke-of-York derivations should be avoided on the grounds of parsimony, but he also claims that there are cases in which a Duke-of-York derivation is either simpler than the alternatives or the only possible explanation. However, McCarthy (2003) argues that the Duke-of-York phenomena described in the literature are vacuous because the intermediate step is not necessary. He claims that non-vacuous Duke-of-York phenomena do not exist in natural language.

It is worth emphasizing that the motivations behind the discussion of cyclic/mutual feeding orders of generative grammars are very different from those of the current article. First and foremost, mal-rules are not intended to behave as a coherent, parsimonious, learnable system of generalizations. Each individual mal-rule represents its own self-contained deviation from a normative grammar. As such, the output of each mal-rule is a spelled out, final wordform. To borrow from the analogy of the Duke of York referenced in Footnote 4, in a generative grammar the Duke of York's intermediate top-of-the-hill is just an inefficient detour. However, in a mal-rule approach to morphological analysis, the top-of-the-hill is an important waypoint along many possible paths an error analysis might take. The problem with mal-rules only arises if a cycle of waypoints leads back to where we came from, because the error tags, which record the path of errors taken, become both self-contradictory and redundant.

Researchers in finite-state morphology have developed a number of solutions to solve the same kinds of problems described by generativists. A two-level morphology (Koskenniemi 1983) compiles all phonological rules into one transducer and can thereby apply all of the rules to the lexicon simultaneously. For example, Karttunen (1993) demonstrates a number of two-level solutions, including one that functions similarly to the disjunctive ordering proposed by Chomsky and Halle (1968). Such applications of two-level morphology are not relevant to the problem of cyclic feeding interactions among mal-rules because they do not allow for keeping and tagging intermediate forms. Individual mal-rules can be implemented as two-level rules, as discussed in Section 2.3 below, but two-level rules cannot solve the problem of cyclic feeding interactions between mal-rules.

2.2. Cycle detection (Graph Theory)

Graph Theory is the study of pairwise relations between objects, typically as part of a network of such relations. We represent feeding relations as directed graphs, where edges go from a given rule's node to the nodes of rules that it feeds into. By using a graph representation, we are able to take advantage of previous research regarding the detection of cycles.

[4]The name comes from the eponymous poem that exhibits a fruitless round trip: *Oh, the grand old Duke of York, He had ten thousand men; He marched them up to the top of the hill, And he marched them down again.*

A cycle in a directed graph can be thought of as an infinite loop, where a given path leads back to its own beginning node. More formally, a directed cycle is a non-empty path in which the only repeated nodes are the first and last nodes, i.e. the first and last nodes in the path are the same node. When representing replace rule interactions as a directed graph, cyclic feeding interactions are conveniently manifested as graph cycles.

Although more efficient algorithms have been put forward, cycles in a directed graph can be detected using depth-first search. If a search finds an edge that points to an ancestor of the current node, the edge to that ancestor is called a *back edge*. All the back edges which depth-first search skips over are part of cycles. Several algorithms have been suggested that are more efficient than naive depth-first search, and we use the algorithm described in Johnson (1975), as implemented in the python package `networkx`[5] (Hagberg et al. 2008).

2.3. Mal-rules

Mal-rules are rules—orthographic, phonological, syntactic, etc.—that generate or license learner errors (Sleeman 1982, Matthews 1992, Antonsen 2012, Reynolds et al. 2022). In the domain of finite-state morphological analysis, mal-rules can be implemented in a number of ways. One example can be taken from Antonsen (2012), who implements mal-rules as part of a two-level ruleset in the following way. First, error tags are added to a path in the lexicon (`lexc`) on both the upper and lower sides. Then, the tag is removed from the lower side under specific conditions using `twolc` rules. The analyses with the error tag in both levels are then removed from the transducer by means of XFST-style regular expression rules. One drawback of this approach is that it does not always allow for combining multiple errors on the same wordform, without significantly complicating interactions between `twolc` rules.

The approach assumed in this article is taken from Reynolds et al. (2022), whose mal-rules are implemented as regular expression replace rules. Figure 1 illustrates this process. Importantly, this entire process is based on a pre-existing transducer that contains the lexicon with all normative surface forms of the target language, along with their morphosyntactic tags. The pre-existing transducer is used to initialize the `main` transducer discussed below.

A replace rule with the optional replacement operator (i.e. `(->)` or `(<-)`) is applied to the main transducer (`main`) to yield an intermediate transducer (`inter1`) which contains everything from the `main` transducer, as well as all forms generated by the rule.[6] The `main` transducer is then subtracted from `inter1` to yield an intermediate transducer (`inter2`) with only those forms that were generated by the rule. Another rule is applied to add an error tag to all readings in `inter2` to yield the last intermediate transducer (`inter3`) which has only the forms generated by the mal-rule, with error tags. The `main` transducer is then replaced by a disjunction of itself with `inter3`, and other mal-rules can then be applied in the same fashion. In this way, mal-rules naturally stack on one another to yield forms that are the combination of multiple errors.[7]

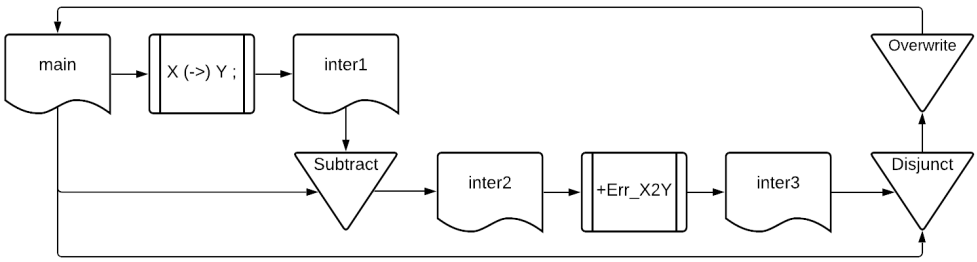


Figure 1: Workflow for adding error(s) to `main` transducer using regular expression replace rules

[5] https://networkx.org/

[6] If a wordform from the original lexicon or preceding mal-rules contains multiple instances of the letter(s) to be changed, the optional replacement operator results in multiple outputs, including every combination of optional changes. In our approach, all of these resulting wordforms are tagged identically, with only one error tag.

[7] It is possible to use `twolc` to achieve a similar result, where instead of applying a replace rule to a fully-compiled transducer, you apply a modified form of the "normative" `twolc` rule set (with only the changes necessary to produce one type of error) to the lexical transducer from `lexc`. Although this approach does give rules access to the underlying forms from the `lexc` transducer, it does not naturally allow for error stacking, since it is based on the output of `lexc`.

3. Methods: Blocking cyclic feeding interactions

While discussing our algorithm for blocking cyclic feeding interactions from our mal-rule application order and maximizing feeding interactions, we use the toy rules in (4) to illustrate the rule interactions.

(4) a. a simple acyclic feeding order (`i → j → k`)
b. a two-rule cyclic feeding order (`m ↔ n`)
c. a 3+-rule cyclic feeding order (`a → e → i → o → u → a`)

3.1. Generating feeding graph from regex rules

The first step of our algorithm is to convert our replace rules into a directed graph, with each node representing a rule, and each directed edge representing a feeding relation from one rule to another. Our method assumes that the replace rules are written as XFST regular expressions, one file per error type. In simple cases, each file may contain a single replacement, and in more complex cases, it may contain multiple parallel replacements or the composition of multiple replacements.

The replacement rule(s) for each error are extracted by reading each source file directly using a simple regular expression to identify every instance of a token followed by one of the optional replacement operators (`(->)` or `(<-)`), followed by another token. Currently, the contextual constraints of conditional replacement (e.g. the `|| L _ R` in the rule `A (->) B || L _ R`) are ignored.[8] This means that the algorithm identifies all true feeding relations (perfect recall), but in cases where constraints block rule interaction, the algorithm generates false positives (imperfect precision). Although we do not yet have a way to automatically detect which constraints block feeding interactions, our script can be run in interactive mode, where a human can manually override these false positives.

The graph is generated by creating a node for each error. Then, for each replace rule belonging to that error, a directed edge is added from that error to any error whose rules take as input that rule's output.[9] An example based on (4) is given in Figure 2. Note that in more complicated real-world examples with multiple replacements per error, there may be more than one edge leaving a node.

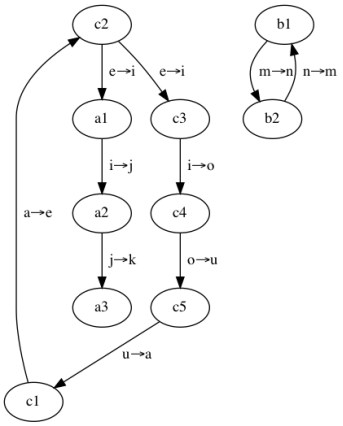


Figure 2: Directed graph generated from rules in (4)

3.2. Blocking cycles

Simple cycles in the rule graph can combine with other cycles to form more complex cycles. As discussed below, our algorithm eliminates cycles by removing all edges between two given nodes, and since removing simple cycles consequently removes their more complex derivatives, our algorithm can focus on removing only simple cycles in the rule graph. To do this, our algorithm starts by addressing cycles of shortest length first, then iteratively removing cycles of greater and greater length until no cycles remain.

As discussed in Section 2.3, we assume an approach that adds errors serially, one after the other. Traditionally, feeding interactions are blocked by putting rules in a counterfeeding order, but this solution does not work for two-rule cyclic feeding interactions.

3.2.1. Blocking two-rule cycles

Two-rule cyclic feeding interactions (a.k.a. "mutual feeding interactions") cannot be placed in a counterfeeding order. Instead, they are added in parallel so that neither rule can feed into the other, as shown in Figure 3. In this figure, both

[8] One reason for this is that transducer composition is non-commutative, so compositions of rule transducers are themselves dependent on the rule ordering. This means that potentially every possible permutation of the rules must be tested to determine whether the rules have a cyclic feeding interaction. This is left to future research, as discussed in Section 5.1.

[9] Errors that feed themselves are assumed to be false positives.

errors are based on the same version of `main`, and their outputs are added back into `main` at the same time, so they can no longer interact. In terms of the rule graph, when two rules are added in parallel, all edges between those two nodes are removed.

In the case of our toy example, the errors associated with rules `b1` and `b2` (`m` → `n` and `n` → `m`, respectively) would be added in parallel.

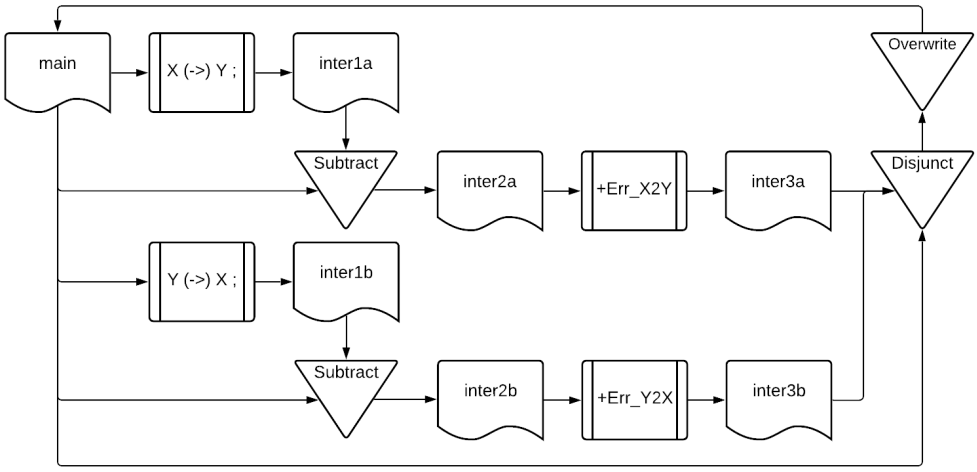


Figure 3: Workflow for adding multiple errors in parallel to `main` transducer using regular expression replace rules

3.2.2. Blocking cycles with more than two rules

In our toy example, the vowel cycle (`a` → `e` → `i` → `o` → `u` → `a`) represents a cycle of more than 2 rules. Cycles composed of more than 2 rules can be be broken by removing only one feeding interaction from the cycle. This can be achieved by putting two rules in a counterfeeding order. For example, if `e` → `i` were ordered before `a` → `e`, then they would be in a counterfeeding order, their feeding interaction would not be manifested, and the cycle would be broken. This is ideal for scenarios in which the optimal ordering described in Section 3.3 below naturally places any two of these rules in a counterfeeding order.

Another approach is to simply add the rules in parallel, as is done with two-rule cyclic feeding interactions. Depending on what parallelized rule sets already exist, parallelizing yet another pair of errors could have unwanted consequences, such as merging two existing parallelized rule sets. For example, if two rule sets already exist, `{A, B}` and `{C, D}`, then parallelizing `A` and `C` would result in merging the two sets into one, which would make it impossible for these errors to stack. The larger the existing rule sets, the greater the impact on stacking.

This scenario poses a dilemma which can only be resolved arbitrarily. Either merge the rule sets to add a large number of errors in parallel (which limits error stacking), or force a counterfeeding order to override the optimal feeding order discussed below is Section 3.3 (which also limits stacking). Here the user should weigh the relative impacts of these approaches on stacking.

Since we already use parallel addition for avoiding two-rule cycles, we default to the same mechanism for blocking cycles with more than two rules. The algorithm can randomly select which two rules to add in parallel, but in interactive mode, the user can manually select the rules (either to add in parallel or to force into a counterfeeding order). In theory, adding any two rules from the cycle in parallel would break the cycle, but in order to maximize the remaining feeding interactions (in the same spirit as Section 3.3 below) we limit the options to removing rules that are *adjacent* in the cycle.

3.3. Methods: Maximizing feeding order

As long as cyclic feeding interactions are effectively blocked, it is usually advantageous to maximize the feeding order of mal-rules so that all possible rule interactions are modeled. In order to estimate the order of rules that maximizes the feeding order, we leverage the work of Gansner et al. (1993), Ellson et al. (2002), whose graph visualization work is implemented in the popular `graphviz`[10] utility. We are most interested in the `dot` algorithm, which is used to render hierarchical drawings of directed graphs.

The first pass of the `dot` algorithm determines the optimal rank assignment of each node consistent with its edges. Roughly speaking, this means that nodes with the most incoming edges are placed at the bottom of the image and nodes with the fewest incoming edges are placed at the top. The result is that the overall flow of the graph runs from top to bottom. All nodes that share the same rank are assigned the same y-coordinates, so the y-coordinate in the output of the `dot` algorithm can be interpreted as a proxy for its rank. Therefore, ordering the mal-rules according to their y-coordinates from top to bottom will optimize their feeding interactions, since the top-most rules have the fewest incoming edges.

The final rule ordering and grouping output by the algorithm takes the order from the `dot` algorithm, and integrates each of the parallelized rule sets by replacing the first instance of one of its members with the entire set. For example, if the rank order from the `dot` algorithm for our toy example were `[b1, b2, c2, a1, c3, a2, c4, a3, c5, c1]` and the parallelized rule sets were `[{b1, b2}, {c1, c2}]`, then the final ordering/grouping would be `[{b1, b2}, {c1, c2}, a1, c3, a2, c4, a3, c5]`.

4. Example from Russian

In order to show a practical example of the algorithm described in Section 3, we apply it to those mal-rules described in Reynolds et al. (2022) that are implemented as regular expression replace rules. A summary of these rules is given in Table 2.

Tag	Tag explanation	Example (Correct form in parentheses)
a2o	Misspelling (о should be а)	озночает (означает)
e2je	Misspelling (е should be э)	ето (это)
Gem	Should be just single, not geminate, letter	рассширить (расширить)
H2S	Misspelling (ь should be ъ)	подьезд (подъезд)
i2j	Misspelling (й should be и)	миллйард (миллиард)
i2y	Misspelling (ы should be и)	блызко (близко)
Ikn	Ikanje (и should be е/я/а)	дитей (детей)
j2i	Misspelling (и should be й)	рабочии (рабочий)
je2e	Misspelling (э should be е)	проэкта (проекта)
NoGem	Geminate letter is missing	имено (именно)
NoSS	Misspelling (ь is missing)	болше (больше)
o2a	Akanje (а should be о)	каторый (который)
prijti	Misspelling the stem of прийти	прийду (приду)
revIkn	Reversed Ikanje (е, а, я should be и)	умерает (умирает)
sh2shch	Misspelling (щ should be ш)	лучще (лучше)
shch2sh	Misspelling (ш should be щ)	вообше (вообще)
ski	по-∼ский instead of по-∼ски	по-русский (по-русски)
SRc	Spelling Rule ы>и (after ц)	близнеци (близнецы)
SRy	Spelling Rule ы>и	книгы (книги)
y2i	Misspelling (и should be ы)	описивают (описывают)

Table 2: Russian mal-rules implemented as regular expression replace rules in Reynolds et al. (2022)

The graph generated from the feeding interactions of these mal-rules has 19 nodes and 139 edges, with 1347

[10] https://graphviz.org

cycles. A visualization of the graph, generated by the `dot` algorithm of `graphviz`, is shown in Figure 4 at a small scale for general reference, but note that digital versions of the document allow for zooming in.

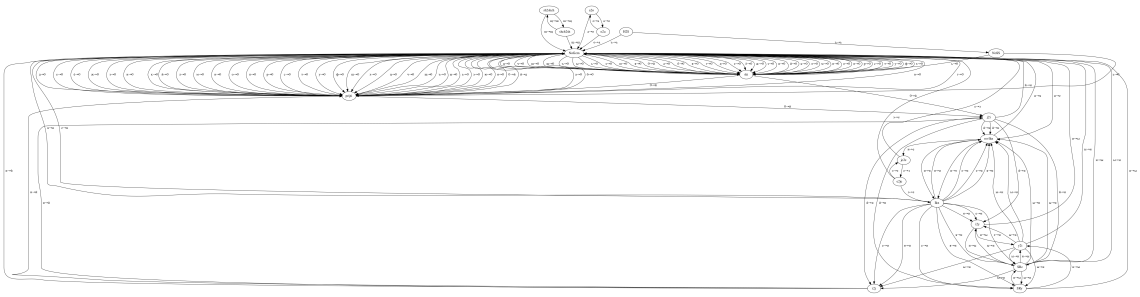


Figure 4: Visualization of the Russian graph based on Table 2 before removing cycles

The algorithm identifies 12 two-rule feeding cycles, and constructs sets of rules that should be added in parallel, as shown in (5). Although most of these pairs are obvious to humans, note that some of the sets have more than two errors because some rules are involved in more than one cyclic feeding interaction.

(5) a. {sh2shch, shch2sh}
b. {a2o, o2a}
c. {Ikn, revIkn}
d. {i2j, j2i}
e. {SRy, y2i, i2y, SRc}
f. {je2e, e2je}
g. {ski, prijti, NoGem}

After removing the 12 two-rule cyclic feeding orders, the remaining graph has 19 nodes and 46 edges, with only one cycle. Note that blocking 12 cycles resulted in the removal of 93 edges. This is because some errors include more than one replace rule that can interact with other rules. For example, the `NoGem` error models when learners fail to write both letters of a geminate, and includes 33 replaces rules, one for each letter of the alphabet.

Also note that although only 12 cycles (i.e., 93 edges) were removed, the total number of cycles in the graph was reduced from 1347 to 1. This is because the majority of the cycles were complex combinations of shorter, simpler cycles. By removing the edges of their component cycles, these complex cycles were also removed.

The only remaining cycle is `prijti` → `j2i` → `SRc` → `i2j` → `prijti`. The algorithm begins by checking whether this cycle is already broken, either because of existing parallelizations, or because the `dot` algorithm currently places the nodes in a counterfeeding order. In this case, the cycle is broken because two of the errors in this cycle are already parallelized: `i2j` and `j2i`. Therefore, no further action is needed to break this cycle.

Although this cycle is already broken, it is worth thinking through what would happen if this were not the case. Although one could randomly select an adjacent pair of nodes to add in parallel, we explore the consequences of this decision. Adding `prijti` and `j2i` in parallel would have the same effect as adding `i2j` and `prijti` in parallel: the sets in (5d) and (5g) would be merged into a set of five rules to be added in parallel. Similarly, blocking either `j2i` → `SRc` or `SRc` → `i2j` would have the same effect regardless: the sets in (5d) and (5e) would be merged into a set of six rules to be added in parallel.

In the spirit of maximizing non-cyclic feeding interactions, adding five errors in parallel could theoretically block fewer feeding interactions than adding six errors in parallel. Otherwise, it could be worth considering which errors in the existing sets are more or less likely to co-occur, preferably on the basis of empirical evidence, and to make the selection to keep co-occuring errors in different sets so that they can stack on top of each other.

In actuality, none of this was necessary, since this cycle was already broken because of pre-existing parallelizations.

Finally, the algorithm plotted the final graph with no feeding cycles using the `dot` algorithm, as shown in Figure 5. Using the y-coordinates as proxy for feeding rank, the algorithm output the following order with parallelized groups

in curly braces: `[H2S, NoSS, {NoGem, prijti, ski}, {e2je, je2e}, {Ikn, revIkn}, {i2j, j2i}, {SRc, SRy, y2i, i2y}, {a2o, o2a}, {shch2sh, sh2shch}]`.

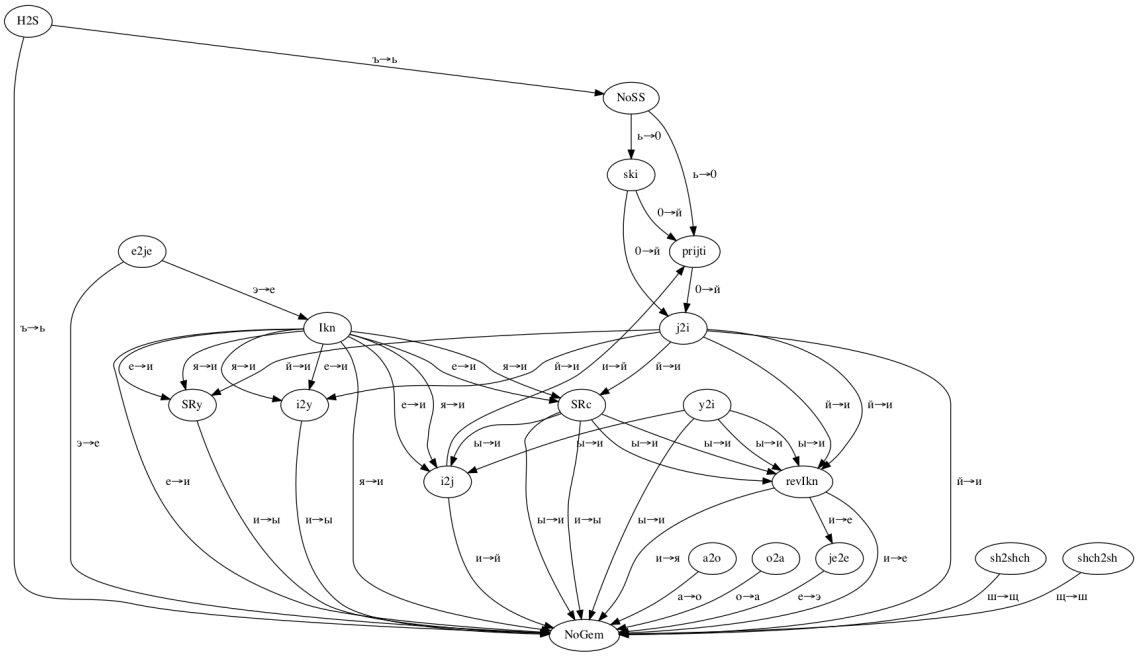


Figure 5: Visualization of the Russian graph based on Table 2 after removing cycles

5. Conclusions

We have discussed the problem of cyclic feeding interactions between mal-rules, and introduced a new algorithm for identifying an ordering (with parallelized groups) to block cyclic feeding interactions and maximize all other feeding interactions. The algorithm can be run in interactive mode, which allows the user to override the automatic behavior of the algorithm.

One unintended benefit of the algorithm is that by making feeding interactions explicit, the user may identify ways to edit replace rules to make them more precise. For example, the `NoGem` rule mentioned in Section 4 was originally written to delete geminate letters for every letter in the alphabet. However, some Russian letters never appear twice in a row, including ь, ъ, and й. By including them in the error rules, unnecessary rule interactions were introduced. The explicit inspection of feeding interactions, as facilitated by our algorithm, can lead to more precise rule writing. Although the intended application of our algorithm is with mal-rules, it could be used for analyzing the feeding interactions of any system of replace rules.

5.1. Future work

The proposed algorithm is focused on *feeding* interactions, but does not consider *bleeding* interactions. Future work can explore how to integrate awareness of bleeding and counterbleeding interactions in the algorithm to maximize error stacking.

As discussed in Section 3.1, our algorithm is not sensitive to conditional constraints on replace rules. The solution to this problem will most likely depend on the existing regular expression compilers for XFST/HFST. For example, one could compile transducers from the replace rules and then compose those transducers independent from any lexicon, and examine the structure of the resulting transducer. However, transducer composition is not a commutative operation,

so the composition of the rules is itself dependent on the order in which the rules are composed. More research is needed to determine whether this (or another) approach could be used to produce a graph of feeding interactions that is sensitive to the conditional constraints of each rule.

Currently, parallelized error sets are added to the final rule ordering at the first occurrence of one of its members in the ranks output by `dot`. Future work is needed to determine how this affects the feeding interactions of other members of the set, and how to optimize the set's position on that basis.

Lastly, the strategy of selecting which errors to parallelize in cyclic feeding interactions of more than two rules is multi-faceted and complex. As discussed in Section 4, this includes the potential merging of existing error sets, as well as the question of maximizing feeding interactions. It is difficult for users to know how a given decision would interact with existing parallelization groups, as well as how it affects future decisions. Future work is needed to help users quickly and easily assess the consequences of parallelizing each pair of errors with respect to these factors.

References

Amaral, Luiz and Detmar Meurers. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL* 23 1: 4–24. https://doi.org/10.1017/S0958344010000261.

Antonsen, Lene. 2012. Improving feedback on l2 misspellings-an fst approach. In *Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*, 080, pp. 1–10. Linköping University Electronic Press.

Choi, Inn-Chull. 2016. Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning* 29 2: 334–364. https://doi.org/10.1080/09588221.2014.960941.

Chomsky, Noam and Morris Halle. 1968. *The sound pattern of English*. Harper & Row.

Chomsky, Noam, Morris Halle, and Fred Lukoff. 1956. On accent and juncture in english. *For Roman Jakobson* 65: 80.

Ellson, John, Emden Gansner, Lefteris Koutsofios, Stephen C. North, and Gordon Woodhull. 2002. Graphviz— open source graph drawing tools. In *Graph Drawing*, edited by Petra Mutzel, Michael Jünger, and Sebastian Leipert, pp. 483–484. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45848-4_57.

Gansner, E.R., E. Koutsofios, S.C. North, and K.-P. Vo. 1993. A technique for drawing directed graphs. *IEEE Transactions on Software Engineering* 19 3: 214–230. https://doi.org/10.1109/32.221135.

Hagberg, Aric, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using networkx. Tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Heift, Trude. 2010. Developing an intelligent language tutor. *CALICO Journal* 27 3: 443–459. https://doi.org/10.1558/cj.27.3.443-459.

Johnson, Donald B. 1975. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing* 4 1: 77–84.

Karttunen, Lauri. 1993. Finite-state constraints. *The last phonological rule* 6: 173–194.

Kiparsky, Paul. 1968. Linguistic universals and linguistic change. In *Universals in linguistic theory*, pp. 170–202. Holt, Rinehart and Winston, New York.

Kiparsky, Paul. 1973. Elsewhere in phonology. In *A Festschrift for Morris Halle*, pp. 93–106. New York: Holt, Rinehart and Winston.

Koskenniemi, Kimmo. 1983. Two-level morphology: A general computational model for word-form recognition and production. Tech. rep., University of Helsinki, Department of General Linguistics.

Matthews, Clive. 1992. Going AI: Foundations of ICALL. *Computer Assisted Language Learning* 5 1: 13–31. https://doi.org/10.1080/0958822920050103.

McCarthy, John J. 2003. Sympathy, cumulativity, and the duke-of-york gambit. *The syllable in optimality theory* pp. 23–76. https://doi.org/10.1017/CBO9780511497926.003.

Meurers, Detmar. 2020. Natural language processing and language learning. In *The Concise Encyclopedia of Applied Linguistics*, edited by Carol A. Chapelle, pp. 817–831. Wiley, Oxford.

Meurers, Detmar, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2019. Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics* 39: 161–188. https://doi.org/10.1017/S0267190519000126.

Nagata, Noriko. 2009. Robo-Sensei's NLP-based error detection and feedback generation. *CALICO Journal* 26 3: 562–579. https://doi.org/10.1558/cj.v26i3.562-579.

Pullum, Geoffrey K. 1976. The duke of york gambit. *Journal of linguistics* 12 1: 83–102. https://doi.org/10.1017/S0022226700004813.

Reynolds, Robert, Laura Janda, and Tore Nesset. 2022. RuMOR: Russian Mentor for Orthographic Rules: ICALL to

help learners of Russian become confident writers. *Computational Linguistics and Text Complexity: a special issue of the Russian Journal of Linguistics* p. 15.

Sleeman, D. 1982. Inferring (mal) rules from pupil's protocols. In *Proceedings of ECAI-82*, pp. 160–164. Orsay, France.

Establishing a Role for Minority Source Language in Multilingual Facilitation

Jack Rueter, Niko Partanen, Khalid Alnajjar and Mika Hämäläinen
University of Helsinki (Rueter, Partanen, Alnajjar, Hämäläinen), University of Turku (Rueter)

Abstract

This document is dedicated to a young man, who, despite the number of times he has traveled around the Sun, is always open to new thoughts on ways to include languages, especially the smaller ones, and the people who speak them in far-reaching and sustainable open-source development. Since Trond Trosterud in Tromsø is attributed a terrific track record in transnational and circum-polar linguistics, we try to attract his attention further afield, to languages and phenomena he has only touched. The language phenomena addressed here come from Erzya and the Zyrian variety of Komi; Erzya has issues presented but not discussed in his dissertation, whereas Komi brings in issues of adnominal and predicate number marking in conjunction with case homonymy that have been resolved thanks to the flexibility of the infrastructure. These source languages, like others, have documented new dimensions and added shape to the ever-growing infrastructure.

Keywords: Erzya, Komi-Zyrian, morphology, multiargument marking

1. Introduction

The idea of establishing an infrastructure that allows work on individual languages of great diversity requires more than a minute for conception. In fact, the formulation of the concept is incomplete without the sleepless nights, nurturing, coaxing, feverish training and jovial interplay with it during its adolescence. It is especially important to note that endangered languages often deal with very different problems than majority languages (Hämäläinen 2021). From a linguistic point of view, infrastructure building requires not only dedication but involvement in the actual study of languages to be addressed on that platform of study as well as an awareness of language needs such as can be observed in earlier categorizations of copula and negation verbs (Trosterud 1994). Another and far reaching language feature subsequently considered in morphological complexity is homonymy that is addressed at an early point (Trosterud 2006) for enhanced application to comprehensive use of the infrastructure.

In 2006, Trond Trosterud published a dissertation on *Homonymy in the Uralic Two-Argument Agreement Paradigms*, where he addressed morphological phenomena beyond those of the Germanic Faroese Trosterud (2009), Saami Antonsen and Trosterud (2011) and Kven Trosterud et al. (2017) languages of Norway. This work was symptomatic of what is required for multilingual facilitation. On the one hand, this work lead to lexical research development, such as discussed in Antonsen et al. (2009b). It also engendered and strengthened concepts of re-usability Antonsen et al. (2010), shared development Gerstenberger et al. (2016); Rueter et al. (2021b). On the other hand, it opened doors to different dimensions of collaboration Trosterud and Moshagen (2021), and even new ones beyond the original infrastructure itself, e.g., Snoek et al. (2014), Simonenko, Alexandra (2020), Alnajja et al. (2020). This also went beyond the principals foreseen by the originators Hämäläinen and Wiechetek (2020), Khanna et al. (2021).

2. Peripheral notes on Erzya

In Trond's dissertation, he addresses several languages, including Erzya and Moksha, whose description we endeavor to expand upon here. Trond's description of four-dimensional paradigms in Mordvin (Trosterud 2006:246-303) gives insight into the phenomena of number and person marking of subject and object on the verb. He provides a plethora of erudite information on the two literary languages and their speakers. Here, we will attempt to put the notes on morphology into perspective, so that in his leisure he might return to this work to update and expand it.

 Nordlyd 46.1: 231-240, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd
https://doi.org/10.7557/12.6370

	ABE	ABL	CMPR	COM	DAT	ELA	GEN	ILL	INE	LAT	LOC	NOM	PRL	TEMP	TRL
PERS PRON	+	+	+	–	+	+	+	+	+	–	–	+	+	–	+
N INDEF	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
N SG DEF	+	+	+	–	+	+	+	–	+	–	–	+	+	–	?
N PL DEF	+	+	+	DIAL	+	+	+	+	+	+	–	+	+	–	+
N PXSG3	+	+	+	–	+	+	+	+	+	–	–	+	+	–	?
ADP	–	+	–	–	–	–	–	–	–	+	+	–	+	–	–

Table 1: cases and locus in Erzya

As noted by Trosterud (2006), Mordvin language morphology deviates from that of the Finnic languages. Whereas verbs in Erzya, Moksha and Finnic take subject marking with distinctions for the categories of person and number, the Mordvin languages also distinguish these two categories in specific subject-object marking on the verb for many instances with singular subjects. A downplayed or third person plural subject in Erzya only distinguishes the person of the object. This resembles singularly object marking structures attested in languages outside the Uralic context, for example Apurinã (cf. Facundes 2000, Rueter et al. 2021a; see also Plains Cree Harrigan et al. 2017). In the nominal system, as in Finnic, the NP head may bear declensional marking specific to the categories of case, number and possessor, but Mordvin includes yet another category – definiteness. All of these categories singularly or in combination might also be followed by a so-called second declension (used for coping with contextual ellipsis), copula person and number marking, which could then be followed by an additive clitic. Hence, in addition to subject-object marking strategies with person and number combinations observed in verbal conjugation, a less frequent phonemon of four-dimensional marking in possessor-index with subject conjugation is also attested in nominals and adpositions, see (1a) below. Furthermore, the NP head in the inessive, abessive, prolative, comparative and translative may also function as adnominal attributes (Rueter 2010:19–22), Erzya and Moksha exhibit so-called secondary declension when the new NP head is lost through contextual/retrievable ellipsis[1]. Perhaps, this will provide for further dimensions to syntax analysis research in Saami language research (cf. Antonsen et al. 2010, Sheyanova and Tyers 2017, Rueter and Hämäläinen 2020a) and beyond (cf. Rueter and Tyers 2018).

Whereas (Trosterud 2006:248) presents a breakdown of the Erzya case systems for number of core, local and other semantic cases with reference to Moksha as well, a rather difficult undertaking, since the two languages have at least slightly divergent polifunctionality in their case systems. For this reason, it might be more to the point of indicating three core cases (NOM, GEN, DAT) and an additional two (ABL, INE), which are also used as object markers (cf. Bartens 1999:91–94, 96; Grebneva 2000:76; Grünthal 2008:229-230, example 6). Likewise, when we speak of basic local cases we must mention a semioverlapping set (ABL, INE, ILL, ELA, LAT, LOC, PRL), which are found in the nominals, postpositions and nouns of deficient declension in the language. The number of cases varies from ten in Gabelentz (1839), and nine in Wiedemann (1865) to fifteen in Rueter (2010). Each enumeration of case is based on different criteria. On the one hand, personal pronouns, which might not exceed nine cases, can be pitted against complex NP heads, which may be found in at least fifteen distinguishable forms, on the other.

In Table (1), NP cases have been dealt with, where possible, using binary + vs – notation. The NP heads serving as locus can be enumerated as follows: PERS PRON 'personal pronouns' (e.g., SG3), N INDEF 'indefinite noun', N SG DEF 'definite singular noun', N PL DEF 'definite plural noun', N PX3SG 'noun with a 3rd person singular possessive suffix', and ADP 'adposition or relational noun'[2]. The pronoun, definite singular and possessive paradigms are smaller than those of the indefinite and definite plural. Some forms, such as the definite plural comitative are only attested in individual dialects (Nad′kin 1968:27–28; Rueter 2010:97–100). Other case forms, such as the translative in *ks* have not been documented in the possessive

[1] Let's say there is a book on the table and another on the shelf. A student takes the *one* on the shelf. In English the word *one* covers for the contextual *book*. In Finnic and Mordvin, however, the noun completely disappears, but even the Finnic requires a verbal form *olevan* 'the one that is' *opiskelija ottaa hyllyllä olevan* (lit. 'student takes on.the.shelf one.that.is'), whereas the NP head morphology locus in Erzya simply shifts to the attribute, i.e., *tonavtŕiićaś saji lavśa langsońťśeńť* (lit. 'the.student takes shelf one.that.is.on') 'the student takes the one that is on the shelf'.

[2] The relational noun *kudikeĺe* 'in the vestibule' (aka postposition or adpostion), although represented by a defective paradigm without a nominative form, can be found in Erzya literature in the definite singular declension, e.g., *kudikeĺgańť* 'vestibule.PRL-DEF.SG' *Kudikeĺgańť tago mařaźevś ĺišića.* 'Once again someone could be heard going out through the vestibule'.

	INDEF	PX1SG	PX2SG	PX3SG	PX1PL	PX2PL	PX3PL
NOM.SG	*jalga*	*jalgam*	*jalgat*	*jalgazo*	*jalganok*	*jalgank*	*jalgast*
NOM.PL	*jalgat*	*jalgan* ∼ *jalgam*	*jalgat*	*jalganzo*	*jalganok*	*jalgank*	*jalgast*
GEN	*jalgań*	*jalgan* ∼ *jalgam*	*jalgat*	*jalganzo*	*jalganok*	*jalgank*	*jalgast*
DAT	*jalgańeń*	*jalgańeń* ∼ *jalgań turtov* ∼ *jalgam turtov*	*jalgaťeń* ∼ *jalgaťturtov* ∼ *jalgat turtov*	*jalgansteń* ∼ *jalganzo turtov*	*jalganok turtov*	*jalgank turtov*	*jalgansteń* ∼ *jalgast turtov*
ABL	*jalgado*	*jalgadon* ∼ *jalgadom*	*jalgadot*	*jalgadonzo*	*jalgadonok*	*jalgadonk*	*jalgadost*
ELA	*jalgasto*	*jalgaston* ∼ *jalgastom*	*jalgastot*	*jalgastonzo*	*jalgastonok*	*jalgastonk*	*jalgastost*

Table 2: the noun *jalga* 'friend'

or definite singular declensions. In a similar vein, the definite plural lative seems to be missing from the enumeration of cases in the Erzya Morphology Grebneva (2000), although it can be attested in the literary works of the prolific Erzya authors Abramov and Shcheglov.

2.1. Adnominal person marking in Erzya

The Erzya language like its sibling Moksha has possessive suffixes as well as verbal and copula conjugation marking, all three of which contribute to the extensive morphologies in the Mordvin languages. Since (Trosterud 2006:264–303) provides ample analyses of verbal conjugations in Moksha and Erzya, there are only certain pieces of information that need to be addressed here, namely possessive person and copula person markers.

The shape of the nominal paradigms can be split into two types. Number is a distinguishing category of the definite declensions, whereas this distinction is only found in the nominative case of the basic or indefinite and possessive declensions. Actually, here the Saami languages documented in Giellatekno describe the abessive case without the category of number, and this idea of zero number can be readily applied to the Mordvin languages for the basic or indefinite declension cases other than the nominative. In the possessive declension, however, the Moksha language actually distinguishes number for three cases, the nominative, genitive and dative, and therefore Erzya has initially been documented with the category for number in these same three cases so as to render symmetric tagging for parallel tool development. The question of whether this should really be necessary is an issue for further development in dialect research Rueter (2020); Rueter et al. (2020a), and shallow-transfer machine translation Rueter and Hämäläinen (2020b), which will also benefit from research at Giellatekno, as alluded to in Trosterud and Antonsen (2020).

In table 2, it will be noted that if there is a distinction made in the form of the possessive marker, then this distinction will be observed in the nominative singular versus other number or case categories. In Erzya, only the third person singular makes this clear distinction in the modern literary language *jalgazo* '(NOM.SG) his/her friend' versus *jalganzo* '(GEN.SG; GEN.PL; NOM.PL) his/her friend's/friends'/friends', whereas the first person singular may become more syncretic following the Southeastern dialect *m*, used in all positions. This information should also be applied to (Trosterud 2006:300–303), where an enhanced understanding of the Erzya paradigm is required, i.e., of the five tables presenting Erzya possessive declension, only table 216, which presents the nominative case, does not require editing.

2.2. Verbal conjugation in Erzya

In an introductory to conjugation Trosterud (2006), Trosterud provides information on the non-past subject conjugation of the verb *śudo|ms* 'to scold' (see table 3), which does not contain two consecutive consonants but does, indeed, retain its stem vowel in the first and second persons plural. Here a vertical line indicates the break between stem and infinitive marker characterized as *-Oms*, where the upper-case *O* is an archivowel, indicative of an obligatory vowel – either the stem vowel or a middle vowel *o*, *e* as assigned by palatal-vowel

	Sing	Plur	Sing	Plur
1	*śudan*	*śudotano*	*-An*	*-Tano*
2	*śudat*	*śudotado*	*-At*	*-Tado*
3	*śudı*	*śudıt́*	*-ı*	*-ıt́*

Table 3: the verb śudo|ms 'to scold' (*not śodoms*) Serebrenikov et al. 1993:635

	Sing	Plur	Sing	Plur
1	*vid́an*	*vid́t́ano*	*pid́an*	*pid́et́ano*
2	*vid́at*	*vid́t́ado*	*pid́at*	*pid́et́ado*
3	*vid́i*	*vid́it́*	*pid́i*	*pid́it́*
	vid́|ems	'to sow'	pid́e|ms	'to cook'

Table 4: the verbs *vid́|ems* 'to sow' and *pid́e|ms* 'to cook' Serebrenikov et al. 1993:133, 476

harmony (Rueter 2010:62–66).

Since then, it has become apparent, upon further scrutiny of the Erzya language, that there actually are near minimal pairs to be found among the verbs *vid́|ems* 'to sow', *vid́ś* 'he/she sowed' and *pid́e|ms* 'to cook', *pid́eś* 'he/she cooked' illustrating single consonants between two vowels. Here, the single consonant *d́* in *vid́|ems* represents the stem final consonant, whereas the segment *-ems* is, in fact, the infinite ending (see Rueter 2016:131).

2.3. Copula complement marking in Erzya

In Trosterud 2006:251–252, the author presents extensive copula paradigms for the consonant-stem noun *sazor* 'little sister' and the mixed stem noun *ovto* 'bear', so it is important that we add the missing vowel-stem noun type *jalga* 'friend', and, instead of paradigms from an older writing tradtion as might be found in (Evsev'ev 1928-29 and Bartens 1999:130–131), we can provide a modern paradigm, which, of course, still requires extensive commenting. The first comment is one to explain ordering of variants in individual paradigm cells, namely, as in our morphological development YAML testing, the first and left-most word form represents the desired form for computer generation.

Vowel-stem nouns never lose their stem-final vowel in the copula declination. The orthography that has developed since the late 1920s, however, has introduced additional syncretism, i.e., whereas the Pre-Soviet standard language (Northwestern dialect) distinguishes the copula form *jalgajan* 'I am a friend' from the possessum form *jalgan* 'my friends', the modern standard embraces either a single form *jalgan* to convey both meanings, or it adds a Southeastern dialect form *jalgam*, which is also syncretic in that it means both 'my friend' and 'my friends'. In the modern Central dialect standard, it must be noted the 1pl form has no final *k* in the non-past regardless of whether it is attached to vowel-stem, consonant-stem or mixed-stem nouns.

	Non-past	Past	Non-past (definite)	Past (definite)
1sg	*jalgan*	*jalgaĺiń*	*jalgaśan*	*jalgaśeĺiń*
2sg	*jalgat*	*jalgaĺit́*	*jalgaśat*	*jalgaśeĺit́*
3sg	*jalga*	*jalgaĺ*	*jalgaś*	*jalgaśeĺ*
1pl	*jalgatano*	*jalgaĺińek* ∼ *jalgatoĺińek*	*jalgaćńet́ano*	*jalgaćńeĺińek*
2pl	*jalgatado*	*jalgaĺid́e* ∼ *jalgatoĺid́e*	*jalgaćńet́ado*	*jalgaćńeĺid́e*
3pl	*jalgat*	*jalgat* ∼ *jalgatoĺ́t*	*jalgaćńe*	*jalgaćńeĺ́t*

Table 5: Copula person and number for the vowel-stem noun *jalga* 'friend'

	Non-past	Past	Non-past (definite)	Past (definite)
1SG	*sazoran*	*sazoroĺiń*	*sazorośan*	*sazorośeĺiń*
2SG	*sazorat*	*sazoroĺiť*	*sazorośat*	*sazorośeĺiť*
3SG	*sazor*	*sazoroĺ*	*sazoroś*	*sazorośeĺ*
1PL	*sazortano*	*sazoroĺińek* ∼ *sazorotoĺińek*	*sazortneťano*	*sazortneĺińek*
2PL	*sazortado*	*sazoroĺiďe* ∼ *sazortoĺiďe*	*sazortneťado*	*sazortneĺiďe*
3PL	*sazort*	*sazoroĺť* ∼ *sazortoĺť*	*sazortne*	*sazortneĺť*

Table 6: Copula person and number for the consonant-stem noun *sazor* 'little sister'

	Non-past	Past	Non-past (definite)	Past (definite)
1SG	*ovtan*	*ovtoĺiń*	*ovtośan*	*ovtośeĺiń*
2SG	*ovtat*	*ovtoĺiť*	*ovtośat*	*ovtośeĺiť*
3SG	*ovto*	*ovtoĺ*	*ovtoś*	*ovtośeĺ*
1PL	*ovtotano* ∼ *ovttano*	*ovtoĺińek* ∼ *ovtotoĺińek* ∼ *ovttoĺińek*	*ovtoťneťano* ∼ *ovttneťano*	*ovtoťneĺińek* ∼ *ovttneĺińek*
2PL	*ovtotado* ∼ *ovttado*	*ovtoĺiďe* ∼ *ovtotoĺiďe* ∼ *ovttoĺiďe*	*ovtoťneťado* ∼ *ovttneťado*	*ovtoťneĺiďe* ∼ *ovttneĺiďe*
3PL	*ovtot* ∼ *ovtt*	*ovtoĺť* ∼ *ovtotoĺť* ∼ *ovttoĺť*	*ovtoťne* ∼ *ovttne*	*ovtoťneĺť* ∼ *ovttneĺť*

Table 7: Copula person and number for the mixed-stem noun *ovto* 'bear'

For many speakers of Erzya, including many grammar writers, the presence of definite noun forms with copula marking is atypical or marginal in the Erzya language (c.f. Evsev'ev 1928-29:125-136, 149-151, 313). The infrequency of these forms in Erzya may, actually, be attributed to the fact that copula person marking occurs on the complement. Thus, both the copula subject and complement must be conceived as highly salient, and the first or second person must be established, i.e., this will not occur in introductions where the proper names are conceivably familiar, but the first and second person pronouns do not indicate a previously establish referent, and therefore are treated as the copula complements.

The mixed-stem noun type, described phonetically in (Rueter 2010:72–73), helps to explain the absence of stem middle vowels in the plural stem of *ovto* 'bear'. In table (7), for example, the second person plural past tense has conceivably three valid forms. The first involves the basic nominative singular form *ovto* form, followed by the past tense marker and second person plural *-ĺiďe*. The second and third forms take the basic nominative plural as their base (*ovtot* and *ovtt*), which are also followed by the past tense marker and second person plural *-ĺiďe*. The distinction, therefore, is that in the mixed-stem type the final middle vowels *o* and *e* (preceded by a soft stem) tend to be dropped before NOM.PL.INDEF, INE.INDEF, ILL.INDEF and ELA.INDEF markers. So far, no instances of word-final *e* preceded by a hard dental have been encountered in this stem type.

All of the examples given in the three tables above (5, 6, 7) address copula person marking on indefinite and definite nouns. Copula person marking can occur with other case marking as well, e.g., it can also occur with the inessive, and both of these cases can, indeed, occur in the possessive declension as well. It will be noted, of course, that nominals do not offer the spate of homonymy that verbs do, but they do present their own variety of four-dimensional marking. Let it suffice to present a nominative singular possessum, with a third person singular possessor and a second person singular copula marker in (1a) and an indefinite inessive possessum with a third person singular possessor, followed by a past tense marker and a first person plural copula marker in (1b). Further information on this can be found in (Gabelentz 1839:237, 402), (Evsev'ev 1928-29:115–125), Turunen (2010) and Rueter (2013).

(1) a. śińďŕe, ton avoĺ trond-oń ćora-z-at??
Sindre, you not Trond-GEN.INDEF son-NOM.PX3SG-COP.NONPST.SC2SG?
'Sindre, aren't you Trond's son?' (p.k.)

	1SG OBJ	1PL OBJ
2SG	*palasamak*	*palasamiź*
3SG	*palasamam*	*palasamiź*
2PL	*palasamiź*	*palasamiź*
3PL	*palasamiź*	*palasamiź*

Table 8: First person indicative object marking for the verb *pala|ms* 'to kiss'

	1SG OBJ	1PL OBJ
2SG	*palamak*	*palamišk*
2PL	*palasamišk*	*palasamišk*

Table 9: First person imperative object marking for the verb *pala|ms* 'to kiss'

b. išak čop kudo-so-nzo-ĺiń
yesterday all.day.long home-INE-PX3SG-COP.PST.SC1SG.
'I was at his/her house all day yesterday' (cf. Evsev'ev 1928-29:62)

As (Trosterud 2006:253–254) continues into the discussion of the subject-object aka definite, objective and object conjugation, he applies the verbal paradigm word *pala|ms* 'to kiss'[3] for the description of subject-object indexing in Erzya, where the syncretism of the *palasamiź* form is stated to render six different readings: Thou kissest[4] us (2SG>1PL), You kiss us (2PL>1PL), You kiss me (2PL>1SG), He/She kisses us (3SG>1PL), They kiss me (3PL>1SG), They kiss us (3PL>1PL), as illustrated in table (8).

The parameters for this paradigm can be defined as follows: (a) a default formative in *-samiź* marks the first person object, (b) the default formative is overridden when both the subject and object arguments are singular by definition. Hence, the formatives *-samak* 'thou kissest me' and *-samam* 'he/she/it kisses me' override the default form. In fact, the six readings given at (Trosterud 2006:254, example 109) might be augmented by an additional interpretation. In the spoken language, the default form is also used when the subject indicates an indefinite but specific actor. This is an aspect that would be important in the development of a rule-based translation machine. If the speaker does not want to reveal that they have been kissed by one person the default form is used. This would be treated as an analogue of the English, where a third person plural form is often used for the same purpose.

In the analysis of the abundance of paradigms afforded in Keresztes (1999) on Trosterud (2006:256–257), the author might take note of a healthy yet critical bit of information from Cygankin (1968:389), such that the formative *-mišk*, in this case *palamišk* 'kiss us!', which is used in the imperative and indicates a default second person subject and a first person object. Only the cell with *palamak* 'kiss me' in it, where both arguments are singular, has an unambiguous reading 1SG>2SG. There should be more villages represented from the Southeast dialects in future research, but for back study, see also (Gabelentz 1839:237, 276), (Wiedemann 1865:74).

As indicated above, all of these notes regarding Erzya and Moksha language morphology and usage are intended to help Trond broaden his horizons in Mordvin language research. But it is definitely the morphophonological and syntactic issues forwarded for each language in this mutual GiellaLT infrastructure that have made us aware of more and more things to write about. And it is this awareness that helps us to transfer comparable questions to other languages in the infrastructure, such as Komi.

[3] first introduced by (Ahlqvist 1859:23-43) in the 1860's to replace the macabre *kalma|ms* 'to bury' verb used by Ornatov (1838) for the illustration of frequentative and habitual verbal derivations

[4] The corresponding second person ending would be *-st*, while *-eth* corresponds to the third person singular *-(e)s* of today's English.

3. Plural Copula Complement Marking in Komi

In the Permic languages, including Komi-Zyrian discussed here, there is a system of plural copula complement marking that differs from the plural marking in other environments. This process is well known and described, but the currently available large Komi corpora allow more nuanced investigation of possible rare features or gaps in the paradigm.

Previously, in discussions of homonymy occuring in illative versus inessive case with possessive marking, it has been noted that constraint grammar rules, such as those discussed in Trosterud (2009), Antonsen et al. (2009c), Antonsen et al. (2009a) must be found for Komi disambiguation. A good example of ambiguity is found in the Komi word for home *горт* and an inessive versus illative singular declension form with third person singular marking *гортас* 'in his/her home' vs 'into his/her home' (see ex. 2a–2b).

(2) a. А сэсся локтӧма горт-ас, Помӧсдінӧ.
and then come.PAST2 home-SG.ILL.PX3SG, Pomøsdin.ILL
'And then he came home, to Pomøsdin.' Пунегова (2021)

b. Бедь кӧ сулалӧ, сідзкӧ, горт-ас некод абу.
stick if stand.PRS.3SG, if.so, home-SG.INE.PX3SG no-one is.not.
'If there is a stick sanding (by the door), that means nobody is at home.' Анохин (2021)

Examples (2a–2b) illustrate distinctions made on the basis of verbal government, such that the verb *локны* 'to come' requires the illative, as seen illustrated in both *гортас* 'into his/her home' and *Помӧсдінӧ* 'to Pomøsdin', while the stationary verb *сулавны* 'to stand' takes locative government, such as the inessive in nouns. These distinctions can also be made by treating the copula complement plural marker in the same way as verbs. In (3a–3b) we can see that the copula plural marker *-ӧсь* is not compatible with an illative reading, thus that reading can be removed in constraint grammar disambiguation. Example (3b), however, provides us with simultaneous noun plural and copula complement plural marking. Hence, we have an analogy for one part of the independent three-dimensional paradigms treated in (Trosterud 2006:164–212), or is this actually four-dimensional, i.e., the possessive suffix introduces the person and number categories associated with the possessor, while the number category of the possessum in *яс* 'pl.' is co-located with the number category of the predicate *ӧсь* 'pl.'.

(3) a. Кирӧ дорын-ӧсь ӧд найӧ, горт-ас-ӧсь.
Kirə at.INE-COP.PL you.know they, home-SG.INE.PX3SG-COP.PL
'But they are ones with Kirø, they're at his home.' Куратова (2020)

b. Найӧ вижов либӧ вежов рӧма-ӧсь, перкальвевъя чут-ъяс-ас-ӧсь.
they yellowish or greenish colored-COP.PL, tan-covered fleck-PL-INE.PX3SG-COP.PL
'They are yellowish or greenish in color with tan flecks.' Ракин (2011)

The ambiguity illustrated in ex (2a–2b) has been one of the challenges in developing Komi-Zyrian Constraint Grammar, and is a classic problem in computational analysis of both Komi literary languages (cf. Rueter et al. 2020b). Examples 3a–3b show, however, that the process and involved questions go even deeper into the plurality and copular constructions. For computational description this is no issue, and the current analyser returns correctly and unambiguously 'горт + Hom1 + N + Sg + Ine + PxSg3 + Pred + Pl' and 'чут + N + Pl + Ine + PxSg3 + Pred + Pl', but from the point of linguistic description this type of structures that contain multiple plural markings have largely been outside the grammatical description tradition, and will need to be addressed in further research. Even for our computational description, the presence of repeated plural tags is a question that may need a more elegant solution. We consider this as a good example where computational and linguistic analysis can enrich one another.

4. Conclusion

Our study started with various notes on Erzya morphological paradigms. We added some important notes into the discussion about Erzya adnominal person marking, and touched briefly the verbal conjugation and

copula complement marking. To complement these observations, we also discussed plural copula complement marking in Komi, which has some behavior which we believe has not been previously discussed extensively enough, or possibly even noticed.

We did not include significantly new results or experiments, but added important novel points into the discussion that has evolved for several decades, and we assume will keep on going.

Acknowledgements

Working with Trond has also meant direct or indirect collaboration with people directly connected to the Giellatekno infrastructure, Divvun and a joint GiellaLT. A few people to mention are Lene Antonsen, Chiara Argese, Tino Didriksen, Børre Gaup, Ciprian Gerstenberger, Ryan Johnson, Sjur Moshagen, Thomas Omma, Flammie A Pirinen, Francis Tyers, Heli Uibo, Kevin Unhammer plus many more.

References

Ahlqvist, August Englebreckt. 1859. *Läran om verbet in mordvinskans mokscha-dialekt*. Helsingfors. Frenckell & Son. Akademisk afhandling, som med den vidtberömda Historisk-filologiska Fackultetens vid Kejserliga Alexanders-Universitetet i Finland samtycke till offentlig granskning framställes af August Engelbreckt Ahlqvist, Hist-Fil. Magister.

Alnajja, Khalid, Mika Hämäläinen, Jack Rueter, and Niko Partanen. 2020. Ve'rdd. narrowing the gap between paper dictionaries, low-resource nlp and community involvement. In *Conference: Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pp. 1–6. https://doi.org/10.18653/v1/2020.coling-demos.1.

Antonsen, L., S. Huhmarniemi, and T. Trosterud. 2009a. Interactive pedagogical programs based on constraint grammar. In *Proceedings of the 17th Nordic Conference of Computational Linguistics. NEALT Proceedings Series 4. 2009*, pp. 10–17.

Antonsen, Lena and Trond Trosterud. 2011. Next to nothing – a cheap South Saami disambiguator. In *Proceedings of the NODALIDA 2011 workshop Constraint Grammar Applications May 11, 2011 Riga, Latvia*, edited by Eckhard Bick, Kristin Hagen, Kaili Müürisep, and Trond Trosterud, vol. 14 of *NEALT Proceedings Series*.

Antonsen, Lene, Ciprian-Virgil Gerstenberger, Sjur Nørstebø Moshagen, and Trond Trosterud. 2009b. Ei intelligent elektronisk ordbok for samisk. In *LexicoNordica 16*, vol. 16, pp. 271–283.

Antonsen, Lene, Saara Huhmarniemi, and Trond Trosterud. 2009c. Constraint grammar in dialogue systems. In *NEALT Proceedings Series 2009*, vol. 8, pp. 13–21.

Antonsen, Lene, Trond Trosterud, and Linda Wiechetek. 2010. Reusing grammatical resources for new languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

Bartens, Raija. 1999. *Mordvalaiskielten rakenne ja kehitys*. Suomalais-Ugrilaisen Seuran Toimituksia 232, Helsinki. Bartens, Raija 1999: Mordvalaiskielten rakenne ja kehitys. Suomalais-Ugrilaisen Seuran Toimituksia 232. Helsinki: Suomalais-Ugrilainen Seura.

Cygankin, D.V. 1968. *Očerki Mordovskix Dialektov tom V*, chap. Opyt klassifikacii èrzânskix govorov mordovskogo prisur'â. Mordovskoe knižnoe izdatel'stvo, Saransk. Naučno-issledovatel'skij institut âzyka, literatury, istorii i èkonomiki pri sovete ministrov mordovskoj ASSR.

Evsev'ev, M.E. 1928-29. *Mordovskaâ grammatika*. Central'noe izdatel'stvo narodov SSSR, Moskva. Èrzân' grammatika.

Facundes, Sidney da S. 2000. *The language of the Apurinã people of Brazil (Arawak)*. SUNY Buffalo. PhD Dissertation.

Gabelentz, Herr Conon von der. 1839. Versuch einer mordwinischen grammatik. In *Zeitschrift für die Kunde des Morgenlandes*, II. 2–3, pp. 235–284, 383–419. Druck und Verlag der Dieterlichschen Buchhandlung, Göttingen. Online: https://github.com/rueter/Erzya-grammar-Gabelentz-1838-39.

Gerstenberger, Ciprian, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology* 4: 29–47. https://doi.org/10.3384/nejlt.2000-1533.1643.

Grebneva, A.M. 2000. *Èrzân' kel', Morfemika, valon' teevema dy morfologiâ*, chap. Padežen' luvos'. Re-

spublikanskoj tipografiâs' «Krasnyj Oktâbr'», Saransk. Vuzon' erzân' dy finnèn'-ugran' kužotnesè student[t]nènen' tonavtnemapel'.

Grünthal, Riho. 2008. *Transitivity in Erzya Mordvin*, pp. 235–246. Uralisztikai tanulmanyok. ELTE„ International.

Harrigan, Atticus G., Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of plains cree verbs. *Morphology* 27 4: 565–598. https://doi.org/10.1007/s11525-017-9315-x.

Hämäläinen, Mika. 2021. Endangered languages are not low-resourced! In *Multilingual Facilitation*. Rootroo Ltd. https://doi.org/10.20944/preprints202104.0113.v1.

Hämäläinen, Mika and Linda Wiechetek. 2020. Morphological Disambiguation of South Sámi with FSTs and Neural Networks. In *Conference: Proceedings of the 1st Joint SLTU and CCURL Workshop*, pp. 36–40.

Keresztes, László. 1999. *Development of Mordvin definite conjugation*. Suomalais-Ugrilaisen Seuran toimituksia, 233. Suomalais-Ugrilainen Seura, Helsinki.

Khanna, Tanmai, Jonathan N. Washington, Francis M. Tyers, Sevilay Bayatlı, Daniel G. Swanson, Tommi A. Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation* 35 4: 475–502. https://doi.org/10.1007/s10590-021-09260-6.

Nad′kin, D. T. 1968. *Očerki mordovskix dialektov*, vol. Tom V, chap. Morfologiâ nižnep′ânskogo dialekta èrzâ-mordovskogo âzyka. Mordovskoe knižnoe izdatel′stvo, Saransk.

Ornatov, Pavel. 1838. *Mordovskaja grammatika / sostavlennaja na narechij mordvy mokshi Pavlom Ornatovym*. V Sinodalnoj tip., Moskva.

Rueter, Jack. 2010. Adnominal Person in the Morphological System of Erzya. In *Suomalais-ugrilaisen seuran toimituksia*, 261. Suomalais-Ugrilainen Seura, Finland.

Rueter, Jack. 2013. *Quantification in Erzya*, pp. 99–122. LINCOM Studies in Language Typology. Lincom GmbH, Germany.

Rueter, Jack. 2020. Korpus nacional'nyx mordovskix âzykov: principy razrabotki i perspektivy funkcionirovaniâ/ dejstviâ. In *Финно-угорские народы в контексте формирования общероссийской гражданской идентичности и меняющейся окружающей среды*, pp. 118–127. Izdatel'skij centr Istoriko-sociologičeskogo instituta, Russia. Conference date: 08-10-2020 through 09-10-2020.

Rueter, Jack, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021a. Apurinã Universal Dependencies treebank. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pp. 28–33. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.americasnlp-1.4.

Rueter, Jack and Mika Hämäläinen. 2020a. FST morphology for the endangered Skolt Sami language. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pp. 250–257. European Language Resources association, Marseille, France. https://aclanthology.org/2020.sltu-1.35.

Rueter, Jack and Mika Hämäläinen. 2020b. *Prerequisites For Shallow-Transfer Machine Translation Of Mordvin Languages: Language Documentation With A Purpose*, pp. 18–29. Iževsk: Institut komp'iuternyx issledovanij, Russian Federation. https://doi.org/0.20944/preprints202104.0131.v1.

Rueter, Jack, Mika Hämäläinen, and Niko Partanen. 2020a. Open-source morphology for endangered mordvinic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pp. 94–100. The Association for Computational Linguistics, United States. https://doi.org/10.18653/v1/2020.nlposs-1.13. Workshop for NLP Open Source Software, NLP-OSS ; Conference date: 19-11-2020 Through 19-11-2020.

Rueter, Jack, Niko Partanen, Mika Hämäläinen, and Trond Trosterud. 2021b. Overview of open-source morphology development for the komi-zyrian language: Past and future. In *Proceedings of The Seventh International Workshop on Computational Linguistics of Uralic Languages*, pp. 62–72.

Rueter, Jack, Niko Partanen, and Larisa Ponomareva. 2020b. On the questions in developing computational infrastructure for Komi-Permyak. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pp. 15–25. https://doi.org/10.18653/v1/2020.iwclul-1.3.

Rueter, Jack Michael. 2016. *Towards a systematic characterization of dialect variation in the Erzya-speaking world: Isoglosses and their reflexes attested in and around the Dubyonki Raion*, pp. 109–148. No. 10 in Uralica Helsingiensia. University of Helsinki, Finland.

Rueter, Jack Michael and Francis M. Tyers. 2018. Towards an open-source universal-dependency tree-

bank for Erzya. In *International Workshop for Computational Linguistics of Uralic Languages, IWCLUL*. https://doi.org/10.18653/v1/W18-0210.

Serebrenikov, B. A., R.N. Buzakova, and M.V. Mosin. 1993. *Èrzânsko-russkij slovar'*. Russkij Âzyk, Digora, Moskva.

Sheyanova, Mariya and Francis M. Tyers. 2017. Annotation schemes in North Sámi dependency parsing. In *Proceedings of the 3rd International Workshop for Computational Linguistics of Uralic Languages*, pp. 66–75.

Simonenko, Alexandra. 2020. Existential possession in Meadow Mari. In *Approaches to predicative possession: the view from Slavic and Finno-Ugric*, edited by Dalmi, Gréte and Witkos, Jacek and Ceglowski, Piotr, pp. 162–181. Bloomsbury Academic. https://doi.org/10.5040/9781350062498.ch-008.

Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of plains cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 34–42. Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-2205.

Trosterud, Reino Sindre, Trond Trosterud, Anna-Kaisa Räisänen, Leena Niiranen, Mervi Haavisto, and Kaisa Maliniemi. 2017. A morphological analyser for kven. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pp. 76–88. Association for Computational Linguistics, St. Petersburg, Russia. https://doi.org/10.18653/v1/W17-0608. Online: https://aclanthology.org/W17-0608.

Trosterud, Trond. 1994. Auxiliaries, negative verbs and word order in the sami andfinnic languages. In *Minor Uralic Languages: Structure and Development*, edited by Ago Künnap, pp. 173–181. Tartu.

Trosterud, Trond. 2006. *Homonymy in the Uralic Two-Argument Agreement Paradigms*. Suomalais-Ugrilaisen Seuran Toimituksia 251. Suomalais-Ugrilainen Seura, Helsinki.

Trosterud, Trond. 2009. A constraint grammar for Faroese. In *Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing*, edited by Eckhard Bick, Kristin Hagen, Kaili Müürisep, and Trond Trosterud, vol. 8 of *NEALT Proceedings Series*, pp. 1–7. Northern European Association for Language Technology (NEALT, http://omilia.uio.no/nealt). Electronically published at Tartu University Library (Estonia) http://hdl.handle.net/10062/14180.

Trosterud, Trond and Lene Antonsen. 2020. Hva er viktig for forståelse? Om maskinoversetting fra nordsamisk. In *Bauta: Janne Bondi Johannessen in memoriam*, edited by K. Hagen, A. Hjelde, K. Stjernholm, and Ø. A. Vangsnes, vol. 11(2) of *Oslo Studies in Language*, pp. 489–502. Oslo: University of Oslo. https://doi.org/10.5617/osla.8514.

Trosterud, Trond and Sjur Moshagen. 2021. Soft on errors? The correcting mechanism of a Skolt Sami speller. In *Multilingual Facilitation*, pp. 197–207. https://doi.org/10.31885/9789515150257.19.

Turunen, Rigina. 2010. *Nonverbal predication in Erzya: Studies on morphosyntactic variation and part of speech distinctions*. Ph.D. thesis, University of Helsinki.

Wiedemann, F.J. 1865. *Grammatik der Ersa-Mordwinischer Sprache*, vol. IX №5 of *VII*. Mémoires de l'Académie Impérial des Sciences de St.-Pétersbourg. Online: https://github.com/rueter/Erzya-grammar-Wiedemann-1865.

Анохин, Павел. 2021. Сьöд вöр шöрöд да нюрöд. *Коми му* 2021-08-26.

Куратова, Н. Н. 2020. Марьюшка. *Войвыв кодзув* 1 №2.

Пунегова, Надежда. 2021. Илона Артеева: Уджыд мед вöлі сьöлöм сертиыд. *Коми му* 2021-09-09.

Ракин, А. 2011. Герчкан. *Би кинь* 1 3.

Om kjønn og adjektiv: «Trond er så eksepsjonell/nydelig/skjønn/grønn»

Ingebjørg Tonne[1], Helene Uri og Lars G. Bagøien Johnsen
Universitetet i Oslo, forfatter og Nasjonalbiblioteket

Abstract

Do female and male writers use adjectives differently? This article is a survey of the potentially gendered use of adjectives in Norwegian novels. It is also a tribute to Trond Tosteruds's legendary article on grammatical gender. While our concern here is biological sex of authors and their use of adjectives, the man of the hour was concerned with rule-governed gender on nouns, albeit with a biological accent. Several readers — and listeners at MONS in 1999 — took note of the bold rule that said that oblong objects, not to mention protruding natural formations, are masculine, while natural phenomena such as pits and cavities are feminine (Trosterud 2001).

Keywords: adjective, statistics, gender, litterature

1 Introduksjon

1.1 Kjønn og språkvitenskapen – en kjappis

For de første språkviterne som skrev om kjønn og språk, var utgangspunktet deres eget språk: mannens språk. Kvinnespråk defineres ut fra forskjellene og avvikene – eller manglene sammenlignet med mannens språk. Det mannlige var normen, det kvinnelige måtte spesifiseres: språk og kvinnespråk – omtrent som det fremdeles gjenspeiles i hverdagsspråket: *band* og *jenteband*, *fotball* og *damefotball*, *professorer* og *kvinnelige professorer*. Otto Jespersen (1941) skriver om både språk og kvinnespråk, og kapitlet om dette siste fenomenet står sammen med kapitler om andre språklige avvik og rariteter: barnespråk, utlendingers språk og pidginspråk. Noen år senere, i 1947, kommer *Manns- og kvinnespråk*, en populærvitenskapelig fremstilling skrevet av den norske filologen Georges Abel. Verken Jespersen eller Abel bygger på annet enn egne spredte observasjoner. Det hindrer dem ikke i å komme med nokså friske uttalelser. Ifølge Abel står det dårlig til med kvinnenes språk: mindre ordforråd og enklere setningsbygning. Det vi skal huske når vi, alt etter hvilket kjønn vi innehar, fniser eller humrer av disse og lignende uttalelser, er at de er skrevet i en tid der få kvinner hadde utdannelse og intellektuelt stimulerende yrker.

Robin Lakoffs bok *Language and woman's place* (1975) er en klassiker innen feltet – trass i at boken er introspektiv og ikke-empirisk. Hun startet med denne utgivelsen det som siden er blitt kalt kjønnsforskjellsparadigmet ('the gender differences paradigm'), med en språkbruksteoretisk tilnærming om mannlig dominans, altså at menn bruker språket til å dominere kvinner (Baker 2014). Fishman (1977) mente i forlengelsen av Lakoffs arbeid at kvinner bruker mer krefter på å legge til rette for god kommunikasjon: Hun mente at kvinner gjorde mer interaksjonelt drittarbeid ('interactional shitwork' og 'donkey work'). Senere har mange påpekt at Lakoff overdrev forskjellene mellom kvinnespråk og mannsspråk, slik som for eksempel Cameron (2007): «Forty years after Lakoff's groundbreaking work,

[1] Ingebjørg Tonnes bidrag til dette arbeidet er finansiert av Norges forskningsråd gjennom ordningen Sentre for fremragende forskning, prosjektnummer 223265.

 Nordlyd 46.1: 241–257, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigert av Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publisert ved UiT Norges arktiske universitet.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6386

we've learned that all such generalizations are over-generalizations: none of them are true for every woman in every context (or even most women in most contexts).»

Deborah Tannens arbeider (f.eks. 1992) representerte noe nytt innen forskningen om språk og kjønn. Hun mener at kvinnens språk er et språk som vektlegger og tar utgangspunkt i et fellesskap, mens mannens språk er et språk som vektlegger uavhengighet, hierarki og status. Tannen understreker at menn og kvinner er oppdratt i forskjellige kulturer, og at dette blant annet har resultert i ulike måter å snakke på. Tannen presenterer undersøkelser som underbygger det synet at barn av ulikt kjønn blir behandlet forskjellig, også språklig.

Forskningen om språk og kjønn kan oppsummeres og skissemessig kategoriseres på denne måten (jf. f.eks. Bull 2021):

> **Mangelhypotesen:** Mannens språk er norm, kvinnens språk er avvikende fra normen og derfor per definisjon mangelfullt, slik blant annet Jespersen (1941) fremlegger det.
>
> **Dominanshypotesen:** Ifølge Tove Bull har forskningen på 70- og 80-tallet «eit klart emansipatorisk sikte og såleis eit tydeleg feministisk utgangspunkt» (Bull, 2021, s. 5). Mennenes dominans og makt blir avdekket og betraktet som kvinneundertrykkende. Formålet med forskningen er å påpeke og analysere forskjellene mellom kvinnens og mannens språk. Kvinnespråket er avmaktens språk og en avspeiling av hennes stilling i samfunnet. Hun bruker språket for å kompensere, for å gjøre seg synlig og oppnå en viss form for respekt. Representanter for denne retningen er blant andre Lakoff og Fishman.
>
> **Forskjellshypotesen:** Menn og kvinner vokser opp i og lever i forskjellige kulturer, og derfor er språket deres ulikt. Tannen skriver innenfor denne tradisjonen.

Felles for disse tre forskningstradisjonene er at det er forskjellene mellom kjønnene som vektlegges, selv om forskjellene innad i en kvinnegruppe eller innad i en mannsgruppe vil være store – kanskje like store som mellom de to gruppene.

Begrepet kjønn har endret seg og endrer seg stadig. Det er vanskelig å betrakte kjønn som en uproblematisk, todelt størrelse som menneskeheten deles inn etter, noe man har eller er. Kjønn er noe som blir skapt og gjenskapt, imitert og utfordret kontinuerlig: «Vi skaper oss selv. Språk er et av verktøyene vi bruker for å gjøre dette. I nyere arbeider om språk og kjønn undersøkes også 'fotnotene' og 'unntakene'; man er ikke lenger interessert i bare den typiske heterofile mann og den typiske heterofile kvinne, men også i guttejenta og jentegutten, i transseksuelle, i homoseksuelle. Man undersøker ulike språkpraksiser som etablerer og omskaper kjønnsbegrepene» (Uri 2018).

Vi skal ikke problematisere todelingen ytterligere her. I denne festskriftsammenheng er vi tre jordnære språkvitere som har gjennomført en jordnær, liten test. For oss har det rett og slett handlet om adjektiv og om kjønn i den tradisjonelle, binære forstand – kvinne og mann – for slik var nå engang materialet vi har til rådighet, klassifisert. Med en kvantitativ tilnærming blir det vanskelig å få øye på brudd med eller sjatteringer i binariteten. Men vi vil likevel allerede nå peke på at den metoden vi har valgt, åpner for at en tradisjonell todeling *kan* nyanseres. Det er viktig å huske at kvantitative studier som har som formål å avdekke forskjeller mellom kjønnene, også kan tilsløre forskjellene innad i kjønnskategoriene. Metoden vår bidrar til å vise om en forskjell er persistent innad i kjønnskategoriene – eller ikke. Det gjør vi ved å undersøke om mønsteret i helheten gjenspeiler seg i delene.

Bull (2021) hevder at den kvantitative variasjonslingvistikken er svekket, og at «vi faktisk visste meir om statistiske forskjellar i språket og språkbruken til kvinner og menn, gutar og jenter på 1980-talet enn vi gjer i dag». Vi bidrar med vårt lille arbeid til å gi et svar på de siste tiårenes manglende interesse for kvantitativ forskning og kjønn.

1.2 Kjønn og adjektiv

Jespersen (1941) var overbevist om at «Kvinderne føler langt større trang til at gi deres varme følelser udtryk i sproget, og da de hvert øjeblik kan komme i begejstring selv overfor mange af livets småting, er det let forståeligt, at de kommer til at slide mere på sprogets rosende ord end mændene». Abel (1947) på sin side var overbevist om at kvinnene har en hang, trolig medfødt, til overdrivelser. Det er mulig å forestille seg at både den trangen Jespersen mener å se hos kvinner, og den hangen Abel ser, kan gi seg

utslag i mange adjektiv hos kvinnelige språkbrukere. Vi himler med øynene og fniser (Lars humrer) når vi leser om Jespersens og Abels syn på kvinnespråk, men kan hende skiller ikke synet seg radikalt fra kjønnsstereotype forestillinger i dag. På høstparten 2021 uttalte en 60-årig mann, la oss kalle ham mannen i gata, om kvinners språk at det er «utflytende og beskrivende, vårt språk er mer kortfattet og to-the-point. Mindre dill og pynt og adjektiv».

Lakoff (1975) hevder at kvinner oftere enn menn bruker adjektiv som *charming*, *divine* og *cute* for å uttrykke følelser – det minner jo om Jespersens «varme følelser». Dessuten mener hun at menn (med unntak av homofile og interiørarkitekter) ikke gir uttrykk for et like fininndelt fargespekter som kvinner: «words like beige, ecru, aquamarine, lavender, and so on are unremarkable in a woman's active vocabulary, but absent from that of most men» (ibid, s. 43). Det er heller ikke vanskelig å finne anekdotisk støtte for dette, her fra en kvinne i gata: «Min erfaring er at kvinner generelt har et mye større fargevokabular enn menn. Vi snakker om akvamarin, kobolt, lavendel, oker, terrakotta, lime, oliven, plomme, krem, muldvarp og røkfarget» (kvinne i Facebook-gruppen Språkspalta, sitert fra Uri 2018). Men ved nærmere ettertanke vil nok de fleste tenke som denne kvinnen, som da hun bad sin mann om å kjøpe turkise stearinlys, fikk rosa lys (ibid): «Men jeg kan love at han kan fargen turkis nå! Når det er sagt, kan faren min navnet på flere fargenyanser enn noen andre jeg kjenner. Dette har vel mest med interesser å gjøre? Far driver med rosemaling.»

1.3 Inspirasjon og tidligere studier

I tillegg til Trosteruds filurkattlignende smil under MONS i 1999 er Ben Blatts (2017) bok *Nabokov's favourite word is mauve* vår viktigste inspirasjonskilde. Blatt undersøker engelskspråklig skjønnlitteratur ved å sette sammen store korpus av litterære tekster, og en del av det han gjør, er knyttet til forfatterkjønn. For eksempel finner han at i femti romanklassikere skrevet av mannlige forfattere er det bare seks av romanene som har flere forekomster av *she* enn av *he*. Tilsvarende – og altså motsatt – tall for romaner skrevet av kvinner er 21 bøker: Nesten halvparten av kvinneromanene har flere forekomster av *he* enn av *she*. Og mer deprimerende: Blatt finner det samme mønsteret i nyutgitte, engelskspråklige romaner. Det står ikke noe bedre til i Norge (jf. Uri 2018, Johnsen og Uri 2021). Kjønnet på de fiktive karakterene i romanene spiller òg en rolle. Blatt finner blant annet at enkelte verb særlig brukes når subjektet er kvinne og andre når subjektet er mann. Lignende mønster ble også funnet i norske tekster (Uri 2018, Dyvik 2018) – i norske tekster hulker kvinner mer enn det menn gjør, mens menn på sin side (som Lars ovenfor) humrer betydelig oftere. Blatt ser også på fordeling og frekvens av ord i engelskspråklige romaner ut fra forfatterkjønn. Resultatene hans er ikke overraskende, og fordelingen av ord signaliserer mer enn noe annet tradisjonelle domener, og enda mindre overraskende er det når vi tar inn over oss at mange av romanene Blatt har søkt i, er fra 1800-tallet. Ordene *pillows*, *lace*, *curls* tyder på at romanen trolig er skrevet av en kvinne. Ordene *chief*, *rear*, *civil* peker i retning av en mannlig forfatter. Blatt serverer oss to lister med henholdsvis de ti mest typiske dameordene og de ti mest typiske herreordene. Adjektivfangsten er imidlertid liten – ingen adjektiv hos kvinnenes ti på topp – og adjektivene begrenser seg til disse tre på herrenes liste: *civil*, *bigger*, *public*.

En stor amerikansk undersøkelse av Facebook-oppdateringer (Park m.fl. 2011) viser at kvinner bruker flere positivt ladde adjektiv enn det menn gjør. Det samme viser en liten norsk studie av journalist-språket (Johnsen og Karlsen 2017).

Svendsen (2019) diskuterer undersøkelser som har tatt utgangspunkt i Robin Lakoffs studie, men der undersøkelsene er gjennomført med mer robuste metoder, særlig korpusbaserte. Disse undersøkelsene viser ifølge Svendsen at språkbruksbildet er komplekst og sammensatt når det gjelder hvordan kvinner og menn velger ord, og at det kan være vanskelig å skjelne mellom hva som kan skyldes kjønn, og hva som kan skyldes andre faktorer, som utdanningsnivå, etnisitet og alder (Svendsen 2019).

For engelsk har også Barczewska og Andreasen (2018) ved hjelp av korpusstudier forsøkt å besvare spørsmål som ligner våre, om adjektivbruk og kjønn. Deres undersøkelse er motivert av det de omtaler som uavklarte og selvmotsigende resultater fra tidligere studier: «Although a great number of scholars highlight the differences between women's and men's speech patterns, there are studies which call into question the conclusion of differentiating between female-male styles of communication. [...] Lakoff

(1973; 1975), Brandis and Henderson (1970), Entwisle and Garvey (1972), Hartman (1976), Bamman et al. (2014), Schmid (2003) and Amir et al. (2012) argue that adjective use differs between the sexes, whereas Kr[a]mer (1974) and Cameron (2008) claim that it does not, and Bamman et al. (2014)'s study seems to suggest that not all differences align with the stereotypes as men may be equally, if not more, expressive than women. Baker (2014) is hesitant to come down strongly in support of either position.» (Barczewska og Andreasen 2018, s. 201).

Mangelen på avklaring i tidligere forskning knytter Barczewska og Andreasen i noen grad til det de mener er den tidlige forskningens overdrevne ekstrapolering fra små mengder data og anekdotiske bevis fra muntlig språkbruk. Forskernes intuisjoner, brukt som metode, kan ha gitt de sprikende resultatene, fremholder de, og mener derfor at korpusbaserte studier er det som skal til. De er dermed på linje med Baker (2014), som i sin bok kombinerer korpuslingvistikk og kjønnsanalyse. Og da blir spørsmålet hva slags språklig korpus som brukes (muntlig eller skriftlig, for eksempel) og hvor stort (eller variert) det er. Dessuten er forskningsspørsmålet selvsagt viktig. Det er særlig to typer spørsmål eller tilnærminger som er interessante når det gjelder kjønn og korpusstudier, som Baker (2014) skisserer det. På den ene siden undersøker man kjønn og språkbruk, det vil si at man vil finne ut hvordan de to kjønnene selv bruker språket. På den annen side kan man ta for seg spørsmål om representasjon av kjønn gjennom språk, altså hvordan kjønnene omtales. Det er den første typen tilnærming som er vårt anliggende i denne artikkelen, slik det er for Barczewska og Andreasen som vi nå ser litt nærmere på.

Barczewska og Andreasen bruker korpuset MICASE (Michigan Corpus of Academic Spoken English), et korpus med transkribert tale fra akademiske situasjoner i perioden 1997 til 2002, med ca 1,85 millioner ord. Korpuset har ulike variabler markert, som type interaksjon, om talerne er studenter eller akademisk ansatte, grad av engelskkompetanse og: kjønn. Kvinner står for 54 % av ordmengden i korpuset, mens menn står bak 46 %. Barczewska og Andreasen understreker at de to kjønnene er jevnt fordelt på de ulike situasjonstypene, slik at for eksempel kvinnestemmene ikke først og fremst er fra lunsjpausene mens mannsstemmene foredrar om egne forskningsresultater. Barczewska og Andreasen stiller følgende forskningsspørsmål: Bruker kvinner flere og et bredere utvalg av adjektiv enn menn? De tar for seg åtte adjektiv, som de kaller «basic adjectives»: *good*, *bad*, *big*, *small*, *pretty*, *ugly*, *different*, *important*. Til disse grunnadjektivene legger de en mengde (nær-)synonymer. Her er nettopp synonymene interessante, i og med at de søker å undersøke om kvinner bruker et bredere utvalg adjektiv. (Nær-)synonymene har de funnet i Thesaurus.com. De er, for eksempel for *good*: *acceptable*, *fine*, *excellent*, *exceptional*, *favorable*, *great*, *marvelous*, *positive*, *satisfactory* og *wonderful*, mens det for eksempel for *ugly* er *awful*, *grisly*, *hideous*, *horrid*, *unseemly*, *unsightly*, *disgusting*, *terrible*, *gross* og *unpleasant*.

De finner at kvinnene i korpuset har en høyere adjektivfrekvens enn mennene, og særlig bruker kvinnene flere av de positivt ladde adjektivene enn de negativt ladde (både grunnadjektivene og nær-synonymene). Men menn bruker også flere positivt enn negativt ladde adjektiv. Bruken av adjektivet *marvelous* trekkes fram som overraskende og på tvers av de mer generelle funnene, idet det er mennene som bruker dette mest. På den annen (negative) side finner de at mennene i undersøkelsen bruker adjektivet *wrong* mer enn kvinnene. For øvrig bølger det litt frem og tilbake i funnene: For eksempel bruker kvinner flest nærsynonymer for *big*, mens menn bruker flest nærsynonymer for *small*. I motsetning til hva Lakoff fant, eller snarere mente, finner de at menn bruker *charming* og *lovely* oftere enn kvinner. Konklusjonen til Barczewska og Andreasen er at de finner forskjeller i adjektivbruk hos kvinner og menn, men også at det er en god del likheter. De avslutter slik: «The mixed results of this study, although limited to one genre, do suggest that Lakoff's intuitions, and the stereotypes they represent, do not always hold and more work should be done into the notion of 'feminine' or 'masculine' adjectives.» (s. 211).

Så da griper vi den stafettpinnen (og legger samtidig merke til at dette avlange objektet er et maskulinum).

1.4 Det vi lurer på

Vi søker etter adjektiv i store, norskspråklige skjønnlitterære tekstkorpus, og vi spør: I hvilken grad kan forfatteres bruk av adjektiv kobles til forfatterens kjønn? Bruker kvinner adjektiv oftere enn det menn

gjør? Er adjektivinventaret et annet hos menn enn hos kvinner? Stemmer det at kvinner bruker flere positivt ladde adjektiv enn det menn gjør? Bruker menn færre fargeord enn det kvinner gjør?

I vår undersøkelse speiler vi delvis Barczewska og Andreasens studie av engelsk muntlig språkbruk. Vi tar utgangspunkt i bokmålsoversettelser av seks av de åtte engelske adjektivene fra studien deres, inkludert nærsynonymene. Vi bruker et stort korpus, men vi dykker også ned i mindre delkorpus for å se om adjektivfordelingen i det store korpuset gjentar seg når det brekkes ned.

2 Metode

Nasjonalbiblioteket har digitalisert hele den norske litteraturen og gjort bøkene tilgjengelig for søk. Den skjønnlitterære delen består i skrivende stund av rundt 55 000 bøker[2]. Digitale søk i store mengder bøker kan avsløre mønstre som for alle praktiske formål ellers ville ha vært skjult for oss. Det er stor variasjon individuelt i språkbruk, men når tekstmengden er stor nok, er det mulig å si noe om trender i bruken, og det er også mulig å dykke ned i mengden for å finne eventuell variasjon. Undersøkelsen vår er korpusbasert, som vil si at man bruker korpus for å sjekke ut om visse hypoteser holder i store språklige tekstsamlinger (Baker 2014, s. 15–16).

2.1 Korpusdata

Datagrunnlaget i undersøkelsen er sammensatt: Det dreier seg om et kuratert, høylitterært korpus som består av hundre norskskrevne romaner, der de eldste bøkene er fra 1890-årene og de nyeste fra 2015. Dessuten har vi brukt korpus hentet fra Nasjonalbibliotekets samlinger, valgt ut ved hjelp av tilgjengelige metadata. Vi har plukket ut de bøkene publisert mellom 1960 og 2010 som er klassifisert under deweynummer 839 (skjønnlitteratur), har målform bokmål, og som har informasjon om kjønnet til forfatteren. Resultat er om lag 21 000 bøker. Den klassen av bøker refererer vi til som «totalkorpuset» eller «totalen». Fra totalkorpuset har vi så konstruert, altså «samplet», fem vilkårlige korpus, hver på 2000 skjønnlitterære bøker.[3]

Felles for det kuraterte korpuset, totalkorpuset og de fem samplede korpusene er at halvparten av bøkene er skrevet av kvinnelige forfattere, halvparten av mannlige. De samplede korpusene er fra perioden 1960 til 2010 og er laget ved at vi fem ganger har stukket neven ned i den digitale haugen med bøker skrevet av kvinner og like mange ganger ned i den tilsvarende haugen med bøker skrevet av menn. En håndfull er altså tusen bøker, så de ti håndfullene som utgjør de fem samplede korpusene, består av til sammen 10 000 bøker. Det kuraterte korpuset er bygd opp med romaner skrevet av anerkjente, norske forfattere. I sample-korpusene og totalkorpuset er det mer blandet, en del er oversatt, og mange av bøkene minner om dem vi har stående på hytta: Her er det både romantikk, erotikk og western. Mens det er 16 forlag, samtlige såkalte seriøse, i sving for det kuraterte korpuset, er det 1523 ulike forlag for totalen – flere av dem med utgivelseslister spekket av underlødige titler.

Vi tar ikke ut forfatterinformasjon annet enn kjønn, slik at hver bok i undersøkelsen ikke har noe annen metadata knyttet til seg. Det er ingen gruppering av bøkene utover forfatterens kjønn. Bøkene bidrar i sin tur til helheten med relative frekvenser. Om boken er stor eller liten, vil det kun være proporsjonen av adjektiv den bidrar med i tellingen frem til totalen for et korpus. Et korpus måles med gjennomsnittet av de relative frekvensene.

2.2 Språkdata

Våre seks oversatte adjektiv fra Barczewska og Andreasens studie er *god*, *dårlig*, *stor*, *liten*, *pen* og *stygg*. I det følgende kaller vi disse grunnadjektivene for B&A-adjektivene. B&A-adjektivene er knyttet til følgende seks sett nærsynonymadjektiv som telles opp samlet innad i gruppen:

[2] Det er tall slik de fremkommer gjennom nb.no i mars 2022 ved å velge kategorien skjønnlitteratur på enten bokmål eller nynorsk.

[3] Alt av data og programkode ligger tilgjengelig via Johnsen (2022).

God: akseptabel, fin, utmerket, eksepsjonell, gunstig, flott, fantastisk, positiv, tilfredsstillende
Dårlig: grusom, forferdelig, fryktelig, elendig, uakseptabel, dårlig, feil
Stor: kolossal, betydelig, gigantisk, massiv, enorm
Liten: trang, begrenset, mager, mikroskopisk, beskjeden, kort
Pen: vakker, sjarmerende, elegant, grasiøs, kjekk, nydelig, ryddig, attraktiv, søt
Stygg: upassende, ekkel, grov, ubehagelig

I valget av fargeadjektiver har vi tatt utgangspunkt i Newtons fargesirkel (unntatt det litt mindre frekvente ordet *indigo*): *blå*, *fiolett*, *grønn*, *gul*, *oransje* og *rød*, og så har vi lagt til *hvit* og ellers en del nyanser: *beige*, *burgunderrød*, *grå*, *lilla*, *rosa* og *turkis*.

De adjektivene vi henter ut kvantitative data for, er utvidet med informasjon fra Norsk Ordbank[4] slik at hvert adjektiv bidrar med en liste av fulle former. Det er de fulle formene som er telt mot korpusene og summert opp for hvert enkelt adjektiv. For eksempel vil adjektivet *god* bidra med ordformene 'beste', 'best', 'gode', 'bedre', 'godt' og 'god', og alle bøyningsformer blir med i analysen.

2.3 Eksperimenter

Vi sammenligner totalkorpuset med det kuraterte korpuset og de samplede delkorpusene i tre eksperimenter og vurderer holdbarheten ved å se på variasjonen mellom korpusene. I det første eksperimentet telles B&A-adjektivene med nærsynonymer for henholdsvis kvinner og menn. I eksperiment 2 telles for henholdvis kvinner og menn bare nærsynonymer. I det tredje eksperimentet telles fargeadjektivene fordelt etter kjønn. I tellingen måles adjektivene med relativ frekvens, altså antall adjektiv per ti tusen ord. Informasjonen er tilgjengelig via Nasjonalbibliotekets DH-lab[5].

Om vi finner at et adjektiv brukes mer av kvinner enn av menn i totalkorpuset, ønsker vi å undersøke om det er en faktisk kjønnsforskjell, eller om det for eksempel er noen få forfattere som gjennom en meget omfattende bruk av visse adjektiv gir oss et inntrykk av at det er en kjønnsforskjell. Det undersøker vi ved å splitte opp korpuset i mindre biter for å se om forskjellen fremdeles er til stede i de mindre korpusene. Mens en vanlig statistisk praksis er å resonnere fra del til hele, prøver vi altså her i vår studie å si om de forskjellene som observeres i helheten, også er til stede i delene. Resonnementet blir at dersom forskjellen består på tvers av korpusene, så kan forskjellen attribueres til kjønn.

I tillegg til å sammenligne helheten med delene vil vi se på i hvor stor grad fordelingen av adjektivene i statistisk forstand viser forskjeller. Hvis enkelte adjektiv har større forekomst hos kvinner enn hos menn, kan vi for det første spørre om hvor store de observerte frekvensforskjellene er (statistisk effektstørrelse), men også i hvor stor grad fordelingene er separert fra hverandre, målt i termer av varians eller standardavvik Hver bok gir adjektivet en frekvens.[6] En gruppe bøker, for eksempel et av våre korpus, vil tilordnes et gjennomsnitt av frekvensene for hver bok som inngår i gruppen; samtidig vil det være variasjon innad i gruppen rundt det gjennomsnittet. Den variasjonen måles i standardavvik.[7] Ved å benytte standardavvik kan vi måle i hvor stor grad to grupper av bøker er separert fra hverandre: Separasjon er knyttet til hvor mange standardavvik det er mellom gjennomsnittet av gruppene; jo flere standardavvik, desto høyere separasjon.

Separasjonen som mål er dermed distinkt fra den gjennomsnittlige forskjellen i frekvens mellom observasjonene. Et annet alternativ for separasjon kunne ha vært å benytte en statistisk test og se på såkalte p-verdier, altså sannsynligheter for observasjoner. Men ulempen med dette siste alternativet er at p-verdier er sensitive for gruppenes datastørrelse, slik at samme gjennomsnitt og standardavvik kan gi lave eller høye p-verdier avhengig av hvor mange bøker som inngår i et (del)korpus. I og med at det også

[4] Norsk Ordbank er beskrevet her https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-5/

[5] Les mer om DH-lab ved Nasjonalbiblioteket og dhlab på https://nb.no/dh-lab

[6] Mer presist - en relativ frekvens, altså hvor stor proporsjon adjektivet tar i boken.

[7] Konsulter gjerne en innføring i statistikk for presise definisjoner og formler.

er variasjon i størrelse på korpusene i undersøkelsen, har vi derfor heller valgt å sammenligne dem gjennom gjennomsnitt av frekvenser og standardavvik fremfor gjennomsnitt av frekvenser og p-verdier.

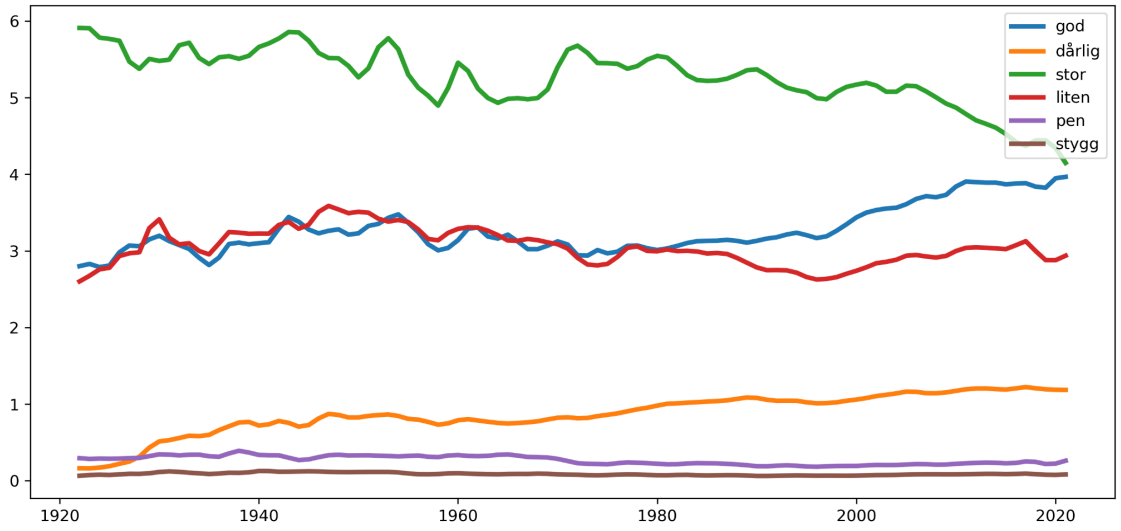
Graden av separasjon presenteres i tre tabeller. De vises som fargekart som letter lesning av mønstre, og for å finne avvikende og lignende tall. Tallene i tabellene er antall standardavvik mellom kjønnene i et korpus. Fargeleggingen er slik at intensiteten følger kolonnen, der adjektivet med størst separasjon vil stå frem med den mørkeste avskygningen. Se for eksempel tabell 2 nedenfor.

3 Resultater

I vår undersøkelse spør vi altså om kvinner bruker flere adjektiv enn menn, og om kvinner bruker andre adjektiv enn menn. Med adjektiv forstås her listene av adjektiv gitt ovenfor (B&A-adjektivene, nærsynonymene og fargeadjektivene). Som nevnt telles B&A-adjektivene med nærsynomymer i eksperiment 1. I eksperiment 2 telles bare nærsynonymer. I det tredje eksperimentet telles fargeadjektivene. Vi gyver løs.

3.1 Eksperiment 1

I dette eksperimentet teller vi B&A-adjektivene og nærsynonymene og sammenligner frekvensen hos mannlige og kvinnelige forfattere. B&A-adjektivene, altså *god*, *dårlig*, *stor*, *liten*, *pen* og *stygg*, har hatt en relativt stabil frekvens i norsk de siste hundre årene som vist i figur 1 nedenfor med en graf fra NB-N-gram som viser trenden for alle bøker uansett sjanger. Selv om trendlinjene for *stor* og *god* endrer seg mest, og tilsynelatende på bekostning av hverandre, er det liten variasjon i frekvens. Eventuelle forskjeller mellom adjektivene kan tilskrives andre variabler. For eksempel er det slik at trendlinjen for adjektivet *dårlig* endrer seg en del fra 1920 til 1930, noe som skyldes etterheng i forbindelse med overgangen fra *aa* til *å* i norsk.

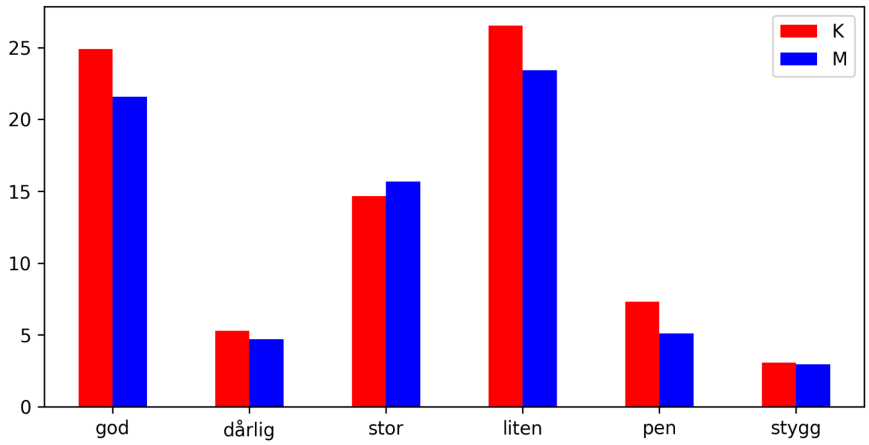


Figur 1: Trendlinjer for adjektivene i studien. Tallene på y-aksen er frekvens pr. 10 000 løpeord.

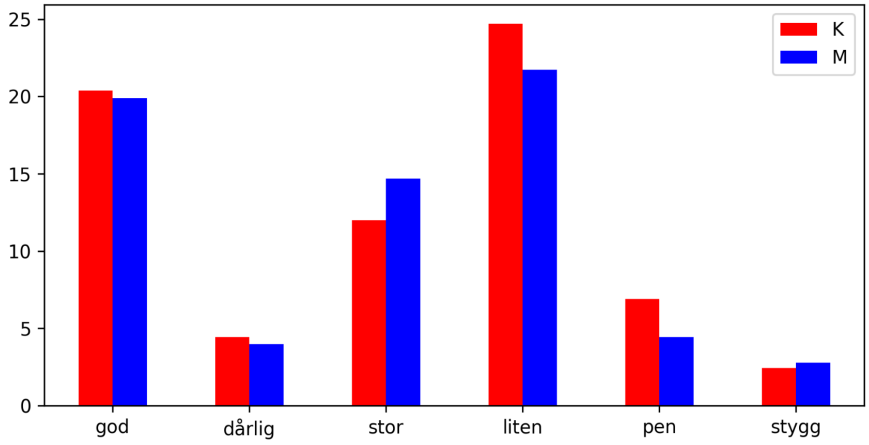
Figur 2 nedenfor viser gjennomsnittlig bruk av adjektiv i totalkorpuset for henholdsvis kvinner og menn, mens figur 3 viser adjektivbruken i det kuraterte korpuset. Her er B&A-adjektivene og nærsynonymene telt, slik at nærsynonymene er innbakt i søylene for sine respektive grunnadjektiv. Alle bøyningsformene telles med. Summen er relativisert til antall ord i teksten de er telt opp i, slik at vi for eksempel ser at alle

formene for *liten* (og nærsynonymene til *liten*) til sammen forekommer ca. 25 ganger[8] pr. 10 000 ord, det vil si omtrent ett adjektiv pr. bokside.

Frekvens av adjektiv fordelt på forfatternes kjønn viser seg å være lik i de to korpusene, slik det fremgår av figurene 2 og 3. De kvinnelige forfatterne bruker B&A-adjektivene (med sine respektive nærsynonymer) noe mer enn menn. Unntakene er adjektivet *stor* som gjennom sine nærsynonymer brukes noe mer av menn enn av kvinner i begge korpusene, og *stygg*, som sammen med nærsynonymene brukes litt mer av menn i det kuraterte korpuset.



Figur 2: Gjennomsnittlig bruk av adjektiv i totalkorpuset, for henholdsvis kvinner (K) og menn (M). Her er B&A-adjektivene og nærsynonymene telt. Y-aksen viser antall adjektiv pr. 10 000 løpeord.



Figur 3: Søylene viser gjennomsnittet for B&A-adjektivene og nærsynonymene, for henholdsvis kvinner (K) og menn (M), i det kuraterte korpuset. Tallene langs y-aksen er antall adjektiv pr 10 000 løpeord.

[8] Økning av den prosentvise andelen sammenlignet med NB-ngram skyldes at i søylediagrammene er alle bøyningsformene både for adjektivet og for nærsynonymene med i opptellingen.

Vi sammenligner i tabell 1 nedenfor også de fem samplede korpusene med hverandre for å undersøke variasjonen av enkeltadjektiv mellom hvert av korpusene. Mens figur 2 gir oss et overblikk over helheten i totalkorpuset, spør vi nå om delene, samplene fra totalkorpuset, oppfører seg likedan. Hvis de gjør det, gir det en indikasjon på om adjektivene er typiske for en kjønnskategori.

For hvert adjektiv måles forskjellen mellom kjønnene som en ratio, der den mest frekvente kategorien divideres på den mindre frekvente. Alle tallene er derfor større enn 1, og jo større tall, desto større forskjell.

	1	2	3	4	5
god	1.2 ♀	1.2 ♀	1.1 ♀	1.2 ♀	1.2 ♀
dårlig	1.2 ♀	1.2 ♀	1.1 ♀	1.1 ♀	1.1 ♀
stor	1.1 ♂	1.0 ♂	1.0 ♂	1.0 ♂	1.0 ♂
liten	1.1 ♀	1.1 ♀	1.1 ♀	1.1 ♀	1.1 ♀
pen	1.6 ♀	1.4 ♀	1.5 ♀	1.5 ♀	1.4 ♀
stygg	1.0 ♀	1.1 ♀	1.0 ♂	1.1 ♀	1.0 ♀

Tabell 1. Hvem bruker adjektiv mest? Tallene viser forskjellen i snitt mellom kvinner og menn målt i ratio, der den mest frekvente kategorien divideres på den mindre frekvente. Alle tallene er derfor større enn 1, og jo større tall, desto større forskjell mellom kjønnene. Røde tall med symbolet ♀ viser til kvinner, blå tall med symbolet ♂ viser til menn. Tallene (1-5) over kolonnene refererer til de fem samplede korpusene.

Resultatene i tabell 1 viser tydelig forskjell mellom kvinner og menn. Det kan være verdt å ha i bakhodet at om det var vilkårlig hvilken kjønnsgruppe som fikk høyest score, vil det være litt over 3 % sjanse for at en hel rad skulle være helt lik. I tabell 1 er hver rad særs lik innad. Legg også merke til raden for *stor*, der frekvensforskjellen er forholdsvis liten[9], men likevel overvekt i mannlig retning i hvert delkorpus. *Stygg* med nærsynonymene *upassende*, *ekkel*, *grov*, *ubehagelig,* som hadde ulik fordeling i totalkorpuset og det kuraterte, bryter mønsteret med ett kjønn over hele linja – bokstavelig talt! Dette er ikke akkurat adjektiv som vi assosierer med jubilanten, men vi vender tilbake til det senere.

I tabell 2 nedenfor gis det første av tre datasett for separasjonsverdier. Tabellen viser hvor mange standardavvik som skiller mellom ♀ og ♂ illustrert i et fargekart, og bør sees i sammenheng med tabell 1 og søylediagrammene i figur 2 og 3. Forskjellen i frekvens (effektstørrelse) kan være stor eller liten og kan variere uavhengig av separasjon, selv om en kan forvente at små forskjeller i frekvens også går sammen med lav separasjonsgrad,[10] altså at om forskjellen er liten den ene veien, kan den lettere gå andre veien, om alt annet er likt. Vi ser for eksempel at separasjonsverdiene for *stor* passer godt sammen med den lille forskjellen i frekvens det er mellom kjønnene for det adjektivet, og at det er det kuraterte korpuset som har høyest grad av separasjon. Samme korrespondanse mellom lav forskjell i frekvens og lav separasjon observeres for det kuraterte korpuset for *god* og *dårlig,* der begge får lav forskjell i frekvens og lav grad av separasjon.

[9] Så liten at forskjellen har forsvunnet i avrundingen.

[10] Det bør sies at alle separasjonsverdiene er lave i og for seg. Med lav forstås her lav sammenlignet med de andre verdiene. En forskjell i standardavvik regnes vanligvis som stor om det er snakk om to eller mer, altså tre til fire ganger så stor som de høye verdiene vi har i våre data.

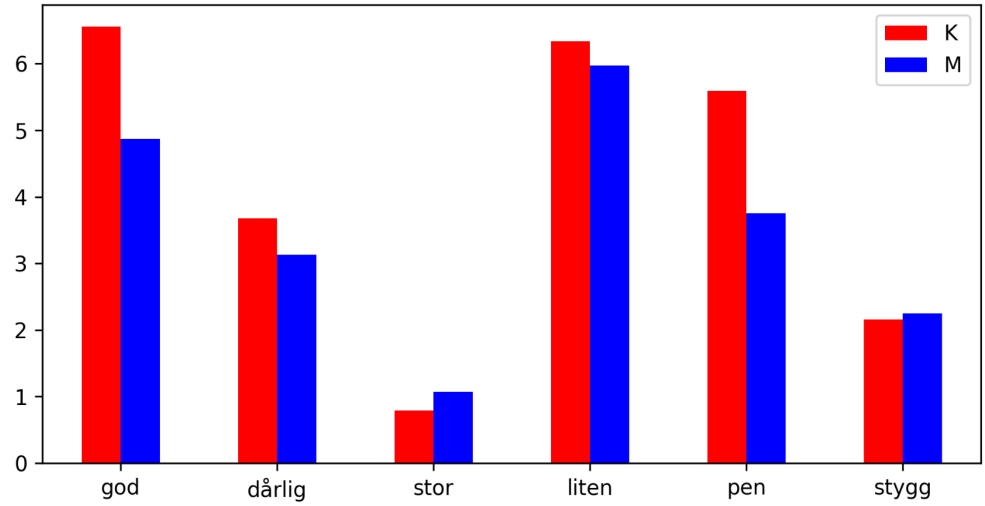
	Total	Kuratert	1	2	3	4	5
god	0,29	0,3	0,31	0,32	0,33	0,34	0,35
dårlig	0,2	0,22	0,24	0,26	0,28	0,3	0,32
stor	0,11	0,4	0,09	0,06	0,04	0,06	0,05
liten	0,21	0,36	0,25	0,22	0,2	0,23	0,23
pen	0,43	0,75	0,51	0,45	0,46	0,51	0,44
stygg	0,06	0,17	0,02	0,09	0	0,06	0,04

Tabell 2: En oversikt over separasjonsverdier i de forskjellige korpusene for S&A-adjektivene med nærsynonymer. Kolonnene indikerer hvilke korpus verdiene kommer fra.

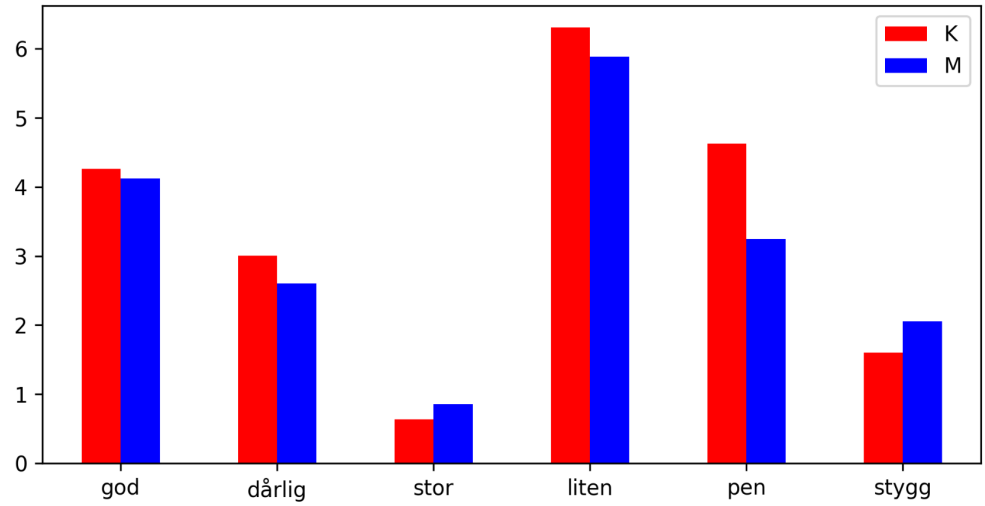
Separasjonstabellene og frekvenstabellene som de fremgår i tabell 1 og 2, samt figur 2 og 3, brukes i resonnementet på den måten at høye verdier forsterker hverandre i retning av en konklusjon om at adjektivene er knyttet til det ene eller andre kjønnet. Fra tabellene og figurene over kan vi konkludere at *pen*, *god*, *dårlig* og *liten* er knyttet til kvinnelige forfattere, og at adjektivet *stor* er knyttet til mannlige: Figurene 2 og 3 viser at *pen*, *god*, *dårlig* og *liten* frekvensmessig er i kvinnenes favør med (relativt) høye separasjonsverdier, mens *stor* går til mannlige forfattere med forholdsvis liten separasjon i total- og delkorpusene, men høy for det kuraterte. Adjektivet *stygg* har vekslende data i sammenligningen og har også gjennomgående lave separasjonsverdier, og peker dermed ikke på noe kjønn. For eksperiment 1 kan vi si at kvinner bruker adjektivene mer enn det menn gjør, unntatt adjektivet *stor* som menn bruker mest, og *stygg* som ikke kan knyttes til noen av kjønnene.

3.2 Eksperiment 2

Som i eksperiment 1 over går vi her gjennom opptellingen i alle korpusene, altså totalkorpuset, det manuelt kuraterte korpuset og det samplede, men nå begrenser vi oss til nærsynonymene og unntar selve grunnformen fra telling. Figur 4 og 5 nedenfor viser resultater for søk i henholdsvis totalkorpuset og det kuraterte korpuset, igjen henholdsvis for kvinner og menn. Søylene er merket med grunnadjektivet, men det er altså nærsynonymene for dette adjektivet som er undersøkt – det betyr for eksempel at søylen merket *pen* viser resultatet av tellingen av *vakker, sjarmerende, elegant, grasiøs, kjekk, nydelig, ryddig, attraktiv, søt.* De nærsynonyme adjektivene forekommer betydelig sjeldnere enn grunnadjektivene.



Figur 4: *Opptelling av nærsynonymer i totalkorpuset for henholdsvis kvinner (K) og menn (M). Antall pr. 10 000 løpeord.*



Figur 5: Opptelling av nærsynonymer i det kuraterte korpuset, for henholdsvis kvinner (K) og menn (M). Antall pr. 10 000 løpeord.

Figur 4 og 5 viser at resultatene av søkene i de to korpusene er ganske like, både når det gjelder bruksfrekvensen for adjektivgruppene og når det gjelder fordelingen mellom kvinner og menn. Sammenlignet med resultatene i eksperiment 1 (jf. figurene 2 og 3) er det en betydelig nedgang for gruppen til *stor*, som har gått fra å være den største gruppen til å bli den minste. Det betyr at forfatterne – av begge kjønn – foretrekker å bruke adjektivet *stor* fremfor nærsynonymene, mens forfatterne for de andre adjektivene bruker både grunnadjektivene og nærsynonymene. Igjen ser vi kvinnenes preferanse for det som kan beskrives som lite, pent og godt. De adjektivene menn har en (svak) tendens til å bruke mer av, sorterer under kategoriene *stygg*, *dårlig* og *stor*.

I tabell 3 under er delkorpusene telt opp og sammenlignet. Den tabellen viser en stor forskjell mellom menn og kvinner for nærsynonymene til *stor*. Nærsynonymene til *stor* – *kolossal, betydelig, gigantisk, massiv, enorm* – brukes betydelig mer av menn enn av kvinner. Adjektivet *stygg* forsterker det som ble antydet i eksperiment 1, altså at menn bruker det mer, altså noe forsterket her (blå, men lave tall hele veien) hvor det bare dreier seg om nærsynonymene – *upassende, ekkel, grov, ubehagelig*. For de resterende adjektivene i hvert delkorpus er det kvinnene som bruker nærsynonymene til grunnadjektivene mest, og kan altså sies å ha en større spredning i adjektivinventaret.

	1	2	3	4	5
god	1.3 ♀	1.4 ♀	1.3 ♀	1.4 ♀	1.3 ♀
dårlig	1.2 ♀	1.3 ♀	1.2 ♀	1.2 ♀	1.2 ♀
stor	1.5 ♂	1.3 ♂	1.3 ♂	1.4 ♂	1.5 ♂
liten	1.1 ♀	1.1 ♀	1.1 ♀	1.1 ♀	1.0 ♀
pen	1.5 ♀	1.4 ♀	1.4 ♀	1.5 ♀	1.4 ♀
stygg	1.0 ♂	1.0 ♂	1.0 ♂	1.1 ♂	1.1 ♂

Tabell 3: Hvem bruker flest av nærsynonymene i hvert av de fem samplede korpusene? Tallene viser forskjellen i snitt mellom kvinner og menn målt i ratio, røde tall med symbolet ♀ viser til kvinner, blå tall med symbolet ♂ viser til menn.

Tabell 4 under viser hvilke separasjonsverdier som korresponderer med data i tabell 3. Det er noen endringer sammenlignet med tabell 2, spesielt ser vi at distribusjonene for *stor* har fått høyere separasjonsverdier, slik at totalkorpus med delkorpus er likt med det kuraterte. Samtidig har separasjonene for *liten* minket.

	Total	Kuratert	1	2	3	4	5
god	0,29	0,3	0,31	0,32	0,33	0,34	0,35
dårlig	0,23	0,24	0,25	0,26	0,27	0,28	0,29
stor	0,28	0,33	0,33	0,24	0,26	0,34	0,34
liten	0,07	0,17	0,08	0,1	0,08	0,1	0,06
pen	0,36	0,59	0,44	0,39	0,39	0,44	0,38
stygg	0	0,25	0,05	0,02	0,03	0,08	0,05

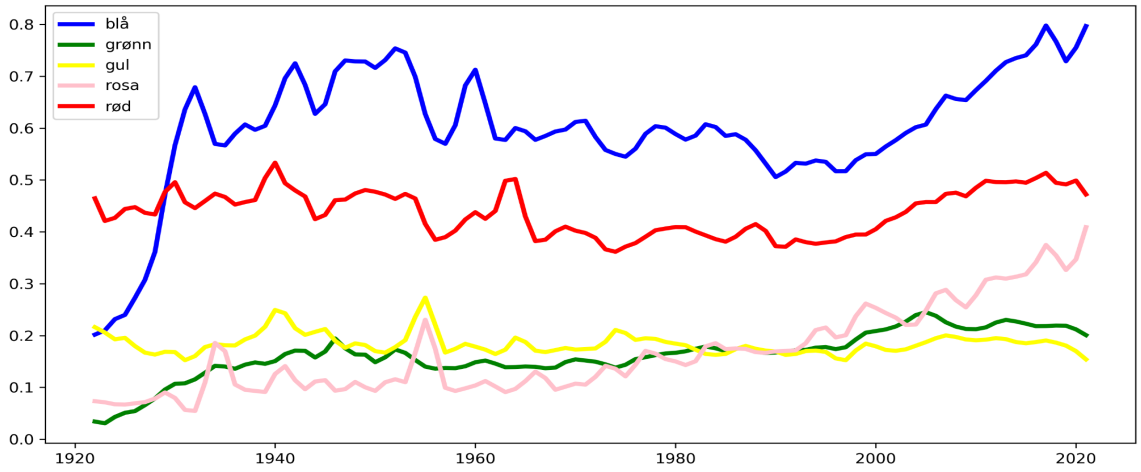
Tabell 4: En oversikt over separasjonsverdier i de forskjellige korpusene for nærsynonymene. Verdiene er utstyrt med en gradert fargestyrke.

Eksperiment 2 viser det samme som eksperiment 1, adjektivene har den samme kjønnstilknytningen som i eksperiment 1. Det er marginale forskjeller i frekvens (mellom kjønnene) for adjektivene, bortsett fra for nærsynonymene til stor der forskjellen har økt en del. Forskjellene med og uten nærsynonymer viser seg i grad av separasjon, kvinner og menn nærmer seg hverandre for *liten*, men går fra hverandre for *stor*.

3.3 Eksperiment 3

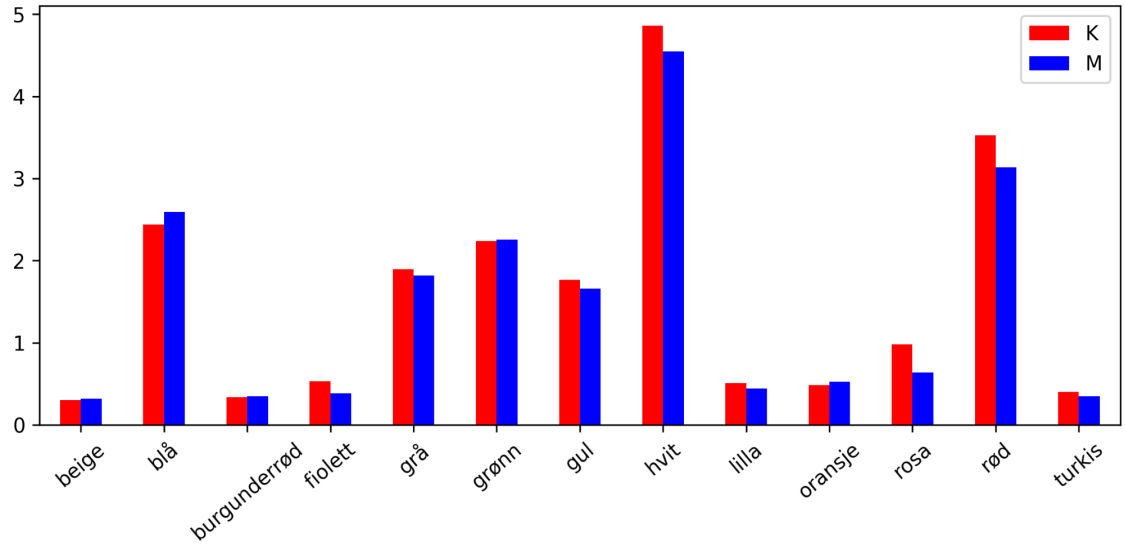
I det tredje eksperimentet teller vi fargeadjektivene for å se om kvinner bruker flere adjektiv enn menn, og hvilke farger som eventuelt foretrekkes av hvilket kjønn.

Bruken av sentrale fargenavn er nokså stabil i norske bøker (figur 6). Vi ser en rosa og gul topp på femtitallet, og det er et lite rosa og blått oppsving på nittitallet.

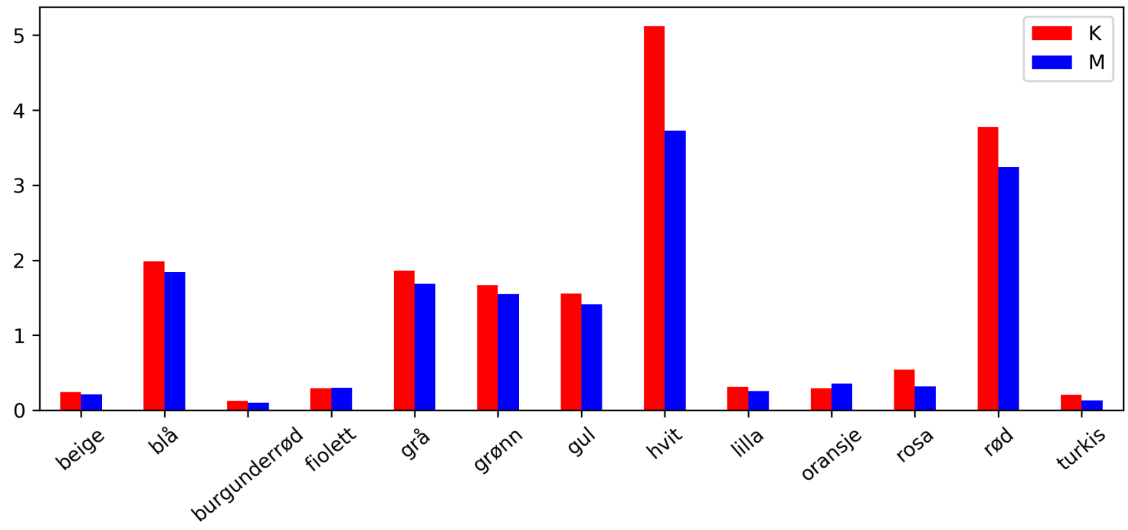


Figur 6: Trendlinjer fra NB-ngram for fem utvalgte farger, vist i antall pr. 10 000 ord.

Figur 7 og 8 under viser distribusjonen i totalkorpuset og i det kuraterte korpuset. Telling av alle våre valgte farger i totalkorpuset er vist i figur 7 og viser stor variasjon mellom fargeadjektivene, det vil si at forholdet mellom menns bruk og kvinners bruk varierer en del fra fargeadjektiv til fargeadjektiv. Figur 8 viser at adjektivet *oransje* har en liten mannlig overvekt i det kuraterte korpuset, og vi ser samme tendens i figur 7 i totalkorpuset. Mens det for B&A-adjektivene (med nærsynonymer) var høy overensstemmelse mellom totalkorpuset og det kuraterte korpuset, er det større variasjon for fargeadjektivene. Noe av variasjonen kan kanskje tilskrives at fargeadjektivene har betydelig lavere frekvens enn B&A-adjektivene med nærsynonymer. Likevel kan vi se et visst mønster.



Figur 7: Oversikt over fargebruken hos kvinner (K) og menn (M) i totalkorpuset.
Tallene langs y-aksen er antall forekomster pr. 10 000 ord.



Figur 8: Fargebruken hos kvinner (K) og menn (M) i det kuraterte korpuset, der y-aksen viser antall forekomster pr. 10 000 ord.

Vi kan fastslå at kvinner er mer fargeglade enn det menn er: Både i figur 7 og 8 ser vi at de fleste fargene er brukt mest av kvinner. I det kuraterte korpuset (figur 8) er 12 av 13 fargeadjektiv hyppigere brukt av de kvinnelige forfatterne; i totalkorpuset (figur 7) er de tilsvarende tallene 10 av 13. Analysen i tabell 5 under viser imidlertid at det er stor variasjon mellom de samplede korpusene med hensyn til hvilket kjønn som bruker de ulike fargeadjektivene mest. Trass i denne variasjonen er det kvinnene som bruker flest fargeadjektiv.

	1	2	3	4	5
beige	1.1 ♂	1.1 ♀	1.4 ♂	1.1 ♀	1.2 ♀
blå	1.0 ♀	1.1 ♂	1.0 ♀	1.2 ♀	1.0 ♀
burgunderrød	1.8 ♀	2.1 ♀	1.4 ♀	1.4 ♂	1.9 ♂
fiolett	1.1 ♀	1.2 ♀	1.0 ♂	1.3 ♂	1.2 ♀
grå	1.1 ♀	1.0 ♂	1.1 ♀	1.1 ♀	1.0 ♂
grønn	1.0 ♂	1.0 ♀	1.0 ♀	1.0 ♀	1.1 ♂
gul	1.1 ♀	1.1 ♀	1.2 ♀	1.2 ♀	1.1 ♀
hvit	1.0 ♂	1.0 ♂	1.1 ♀	1.0 ♀	1.0 ♀
lilla	1.5 ♀	1.1 ♂	1.0 ♂	1.1 ♂	1.1 ♀
oransje	1.5 ♀	1.1 ♀	1.1 ♀	1.1 ♀	1.7 ♀
rosa	1.4 ♀	2.0 ♀	1.4 ♀	1.3 ♀	1.5 ♀
rød	1.1 ♀	1.2 ♀	1.2 ♀	1.3 ♀	1.2 ♀
turkis	1.8 ♀	1.0 ♀	1.4 ♀	1.1 ♂	1.4 ♀

Tabell 5. Gruppevis fordeling av kjønn over farger. Tallene viser forskjellen i snitt mellom kvinner og menn målt i ratio. For eksempel viser tallet 2 i raden for rosa at det adjektivet forekommer over dobbelt så hyppig hos kvinner som hos menn. Røde tall med symbolet ♀ viser til kvinner, blå tall med symbolet ♂ viser til menn.

Tabell 6 under viser oversikten over separasjonsverdier for fargeadjektiv i alle korpusene. På samme måte som over benyttes fargekode for å angi grad av separasjon, sammen med en tallfesting.

	Total	Kuratert	1	2	3	4	5
beige	0,07	0,2	0,03	0,1	0,14	0,09	0,17
blå	0,03	0,11	0,03	0,05	0,01	0,09	0,01
burgunderrød	0,1	0,36	0,32	0,45	0,26	0,12	0,17
fiolett	0,01	0,03	0,05	0,11	0,02	0,12	0,09
grå	0,02	0,14	0,05	0,03	0,11	0,07	0
grønn	0,02	0,08	0	0	0,02	0	0,04
gul	0,04	0,13	0,05	0,04	0,07	0,09	0,04
hvit	0,03	0,4	0,01	0,01	0,06	0,03	0,02
lilla	0,06	0,18	0,21	0,06	0,01	0,05	0,05
oransje	0,09	0,14	0,11	0,12	0,06	0,08	0,29
rosa	0,14	0,56	0,18	0,16	0,17	0,17	0,24
rød	0,15	0,23	0,11	0,17	0,15	0,22	0,17
turkis	0,06	0,69	0,24	0,01	0,15	0,07	0,14

Tabell 6: En oversikt over separasjonsverdier i de forskjellige korpusene for fargeadjektiv. Kolonnene indikerer hvilke korpus verdiene kommer fra.

Den mest typiske kvinnefargen er rosa, som er hyppigst brukt i bøker skrevet av kvinner, og som i tillegg har høy grad av separasjon i alle korpusene. Fargen *rød* knytter seg også tydelig til kvinnelige forfattere, med høy frekvens og høy grad av separasjon.

La oss fra fargefunnene leke oss med en rangering av fargeadjektivbruk og kjønn. Det kan vi gjøre med utgangspunkt i resultatene vi ser i tabellene 5 og 6, samtidig som vi skjeler til tallene i figurene 7 og 8, med henholdsvis totalkorpuset og det kuraterte korpuset. Leken går sånn: Vi ordner fargeadjektivene etter hvor mange «stemmer» de får fra å ha høyest verdi i de forskjellige korpusene (ett korpus = én stemme), og der fargene scorer likt, ordnes de igjen etter en vurdering av frekvens og grad av separasjon. Rangeringen gir toppscore til kvinner for adjektivene *rosa, rød* og *gul*, rangert etter frekvens. Dernest ser vi på de adjektivene som har minst seks kvinnestemmer, altså der kun ett korpus avviker, der finner vi *turkis* og *oransje*, begge med varierende frekvens og separasjon. Blant de som avviker med to stemmer finner vi *hvit*, *grå* og *blå* alle for kvinner, der *hvit* tar plassen foran de andre med størst frekvensforskjell i det kuraterte. Til slutt, om vi fordeler adjektiv som har fire kvinnestemmer og tre mannsstemmer, finner vi *burgunderrød, fiolett, grønn* og *lilla.* Adjektivet *grønn* skiller seg ut med lavest separasjonsgrad, mens *burgunderrød* er interessant ved at det skifter mellom korpusene, og tillegg har høy grad av separasjon, selv når det skifter side. Så topplisten for feminint foretrukne farger er: *rosa, rød, gul, turkis* og *oransje*. På ventelisten setter vi *blå, grå, burgunderrød, hvit* og *turkis*. For menn kunne vi våge oss til å sette opp de to adjektivene *grønn* og *lilla*, som får tre stemmer hver for mest mannebruk, og har mest konsistent separasjon.

4 Mer fargelegging og *Den gode, den onde og den grusomme*[11]

Søkene i eksperiment 1, med utgangspunkt i B&A-adjektivgruppene, viser at kvinner bruker hver adjektivgruppe mer enn det menn gjør. Unntaket er adjektivgruppen for *stor*. Om vi titter inn i hvert av de fem samplede korpusene, ser vi at kvinner bruker adjektiv mer frekvent enn det menn gjør – bortsett fra for adjektivgruppen til *stygg,* som i tre av de fem samplede korpusene brukes mest av menn, og som også har lave separasjonsverdier. For adjektivgruppen til *stygg* finner vi dermed ingen klar kobling til kjønn. I eksperiment 2, der bare nærsynonymene er telt, er kvinnelige forfatteres tendens til å bruke flere adjektiv enn sine mannlige kolleger enda klarere enn i eksperiment 1. For tre av adjektivene – *god, liten, pen* – er tallenes tale ekstra tydelige. Unntakene fra kvinnedominansen er nærsynonymene til *stor*, som brukes mest av menn, og *stygg*, som ikke har noe dominant kjønn. For de resterende adjektivene er det kvinnene som bruker nærsynonymene mest.

Dersom vi later som om verden er inndelt i det som er godt og det som er dårlig, og lett kan beskrives enten med positive adjektiv eller negative adjektiv, ser vi at kvinner bruker positivt ladde adjektiv i enda større grad enn sine mannlige kolleger. Både grunnadjektivene *god* og *pen* og nærsynomene til disse brukes mer frekvent av kvinner. At mennene bruker nærsynonymene til *stygg* mer enn kvinner gjør, peker i retning av at menn i større grad enn kvinner bruker de negativt ladde adjektivene. Men tallene er ikke krystallklare for positiv-negativ ladning: Kvinner bruker *dårlig* og og dennes nærsynonymer oftere enn det menn gjør.

Eksperiment 3 viser at de fleste fargene er brukt mest av kvinner. Likevel er det ikke mange adjektiv som kan sies å være *typiske* for kvinner. For selv om kvinner bruker flere av omtrent alle fargeadjektivene, vil variasjonen i gruppene være såpass stor at få adjektiv peker seg ut som først og fremst kvinnelige – rett og slett fordi mannlige forfattere også bruker dem ganske mye. Det er kanskje bare *rosa* og *rød* vi rimelig trygt kan hevde er typisk kvinnelige farger, mens *blå*, *oransje* og *gul* er de neste kvinnefargekandidatene. Og for menn: Kanskje vi kan driste oss til å hevde at *grønn* – og muligens *lilla* – er maskuline adjektiv. Våre hetero-normative sjeler stusser over at lilla ser ut til å være en mannefarge, men forskning skal jo nettopp flytte forestillinger og mulige fordommer.

For fargeadjektivene er det altså mer kvinnebruk enn mannebruk, men ikke mye mer og ikke for alle fargene. Resultatene kan dermed bare i noen grad sies å stemme med det Lakoff (1975) hevdet om fargeadjektiv. Lakoff brukte ikke korpus, men egne intuisjoner og inntrykk fra bekjente, som minner om hvordan Jespersen (1941), Abel (1947) og vårt 2021-eksemplar av arten mannen i gata gikk til verks.

Oppsummert finner vi at hovedspørsmålet vi har stilt, om kvinner bruker adjektiv mer enn menn i vårt datamateriale, kan besvares med «ja». Når det gjelder spørsmålet om kvinner bruker andre adjektiv enn menn, om de varierer mer, er svaret også «ja», kanskje i retning av «tja», idet kvinner bruker nærsynomymer oftere enn mennene, selv om noen få adjektiv brukes oftere av menn.

Resultatene fra vår undersøkelse viser på mange måter det Barczewska og Andreasen (2018) konkluderer med for sin del – at ulike tilnærminger og metoder vil gi ulike resultater. Vi kan legge til at også ulike korpustyper og adjektivutvalg sannsynligvis også har mye å si for resultatet (jf. også Baker 2014). Vi har brukt korpus med skriftlig tekst, et skjønnlitterært basert korpus. Fordelen med store, skjønnlitterære korpus er at man kan forvente et vell av ulike situasjoner, et hav av mellommenneskelige interaksjonstyper og relasjoner. Ulempen er at forskjeller innad i en gruppe kan utlignes; noen deler av korpus har høye verdier, mens andre deler har lave, og snittet som presenteres, er kanskje litt villedende. Med vår samplingsmetode har vi søkt å gjøre noe med denne ulempen. Samplingen av korpuset viser at det kan være noen individuelle forskjeller hos de ulike forfatterne når det gjelder bruken av adjektiv, men også at det store korpuset er relativt homogent med hensyn til kvinners og menns adjektivbruk. Svendsen (2019), omtalt tidligere, er inne på denne problematikken når hun fremholder: «While all studies discussed in this article take an empirical approach to examine Lakoff's claims about women's language, they still share a general methodological problem: all investigate the differences between the speech of

[11] Italiensk westernfilm fra 1966 – som jubilanten bør se hvis han ikke allerede har gjort det – regissert av Sergio Leone, originaltittel *Il buono, il brutto, il cattivo*.

men and women assuming that men and women form homogenous groups.» Hun mener at det er vanskelig å isolere kjønnseffekt i og med at faktorer som «ethnicity, age, educational level, socio-economic status, sexuality, etc. vary within the groups and may indeed affect the speech style of the individual» (Svendsen, 2019, s. 8). Også vi forholder oss til en todeling av menneskeheten i enten kvinner eller menn, men vi går altså i vår undersøkelse med samplingen også inn i delkorpus og finner da relativt stor grad av homogenitet, med noen punkter med intern heterogenitet.

Samtidig er vårt korpus delt inn etter forfatterens kjønn, ikke kjønnet til den som har synsvinkelen i romanen. Hvis en mannlig forfatter ønsker å skrive ut fra en kvinnelig synsvinkel, hva da med adjektivbruken? Det blir kanskje en annen skål. I mellomtiden skåler vi for jubilanten i alkoholfri vin: Skål! Og hvis vi skulle valgt tre adjektiv som beskrev ham, ville vi verken ha valgt de typiske manneadjektivene *lilla*, *stygg* eller *stor*, men kanskje snarere *rød*? Og dessuten *dyktig* og *sympatisk*.

5 Referanser

Abel, Georges. 1947. *Manns- og kvinnespråk: en populær fremstilling*, Oslo. https://www.nb.no/nbsok/nb/4a41b0e618c97415d4c32872da0a8670?lang=no#0

Amir, Zaini, Hazirah Abidin, Saadiyah Darus, and Kemboja Ismail. 2012. Gender differences in the language use of Malaysian teen bloggers. *GEMA Online™ Journal of Language Studies* 12(1): 105–124.

Baker, Paul. 2014. *Using corpora to analyze gender*. Bloomsbury, London, New York.

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2): 135–160. https://doi.org/10.1111/josl.12080

Barczewska, Shala og Andreasen, Agata. 2018. "Good or marvelous? Pretty, cute or lovely? Male and female adjective use in MICAS". *Suvremena Lingvistika* 44(86):194-213. https://doi.org/10.22210/suvlin.2018.086.02.

Blatt, Ben. 2017. *Nabokov's Favorite Word is Mauve*. New York: Simon & Schuster.

Bull, Tove. 2021. Kjønna språk og språkbruk før og no. *Målbryting* nr.12(2021): 1–24. https://doi.org/10.7557/17.5787.

Cameron, Deborah. 2008. *The Myth of Mars and Venus*. Oxford University Press.

Dyvik, Helge. 2018. https://norgramtall.w.uib.no/om-denne-bloggen/

Fishman, Pamela M. 1977. 'Interactional Shitwork'. *Heresies* 2: 99–101.

Hartman, Maryann. 1976. A descriptive study of the language of men and women born in Maine around 1900 as it reflects the Lakoff hypotheses in "Language and Women's Place". Paper presented at the Conference on the Sociology of the Languages of American Women. New Mexico State University, Las Cruces, New Mexico, January 16–17. Available online at https://files.eric.ed.gov/fulltext/ED130316.pdf.

Jespersen, Otto. 1941. *Sproget. Barnet, kvinden, slægten*. Gyldendalske boghandel.

Johnsen, Lars G. Bagøien. 2022. Datasett https://doi.org/10.5281/zenodo.6425086

Kramer, Cheris. 1974. Women's speech. Separate but unequal? *Quarterly Journal of Speech* 60(1): 14–24. https://doi.org/10.1080/00335637409383203

Lakoff , Robin T. 1975. *Language and woman's place*. New York: Harper & Row.

Schmid, Hans-Jörg 2003. Do women and men really live in different cultures? Evidence from the BNC. In *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech* edited by Andrew Wilson, Paul Rayson, and Tony McEnery, 185–221. Peter Lang.

Svendsen, Amalie Due. 2019. Lakoff and Women's Language: A Critical Overview of the Empirical Evidence for Lakoff's Thesis. *Leviathan: Interdisciplinary Journal in English*, 4, 1-11. https://doi.org/10.7146/lev.v0i4.112651

Tannen, Deborah. 1990. *You just don't understand: Women and men in conversation*. New York: Morrow.

Trosterud, Trond. 2001. Genustilordning i norsk er regelstyrt. *Norsk lingvistisk tidsskrift*, 19: 29–58.

Uri, Helene. 2018. *Hvem sa hva? Kvinner, menn og språk.* Oslo: Gyldendal.

Projections for Sámi in Norway: Schools as Key to Revitalization

Øystein A. Vangsnes
UiT The Arctic University of Norway & Western Norway University of Applied Sciences

Abstract

Based on the contemporary number of children receiving instruction in Sámi in the compulsory part of the Norwegian school system (grades 1–10), the paper presents three different projections of the future number of Sámi language users in Norway. There exist three different curricula for the subject Sámi, one for first language pupils (Sámi 1), one for second language pupils (Sámi 2), and one for pupils with no previous knowledge of the language (Sámi 3). Depending on whether only Sámi 1 pupils become future language users, or also Sámi 2 or even Sámi 3 pupils do so, a sober, moderate, and optimistic prognosis can be made, respectively. The sober prognosis predicts a dramatic decrease for North Sámi, a slight decrease for Lule Sámi, and status quo for South Sámi, whereas the moderate prognosis predicts status quo for North Sámi and an increase for Lule and South Sámi, and the optimistic prognosis predicts an increase for all three varieties, but most notably for Lule and South Sámi. A number of factors that are likely to modulate the prognoses are brought to attention and discussed, unveiling that more information is needed regarding a number of issues that bear on how the future of the Sámi languages in Norway can be estimated.

Keywords: Sámi languages, revitalization, Sámi in Norway, language vitality

1. Introduction

In this paper, estimates of the future population of Sámi language users in Norway are made on the basis of the current number of pupils receiving instruction in one of the Sámi languages. This is rooted in the assumption that literacy is one of the most powerful factors in the language ecology of contemporary Sámi and that the school system is a very efficient tool for advancing literacy and language proficiency in the Sámi languages. Other factors are also important to take into consideration when assessing the vitality of a language, and the school-based prognoses must inevitably be integrated in a broader model where a range of factors modulate the total number of language users.

The Sámi languages are indigenous to the High North of Europe. The traditional homelands of the Sámi are the central and northern parts of Norway and Sweden, the northern part of Finland and the Kola Peninsula in Russia. The North Sámi autonym for this area is Sápmi, and variants of this name are used in other Sámi languages.[1] This paper will be confined to an assessment of Sámi in Norway for which we have exact figures regarding the number of pupils with Sámi language instruction and furthermore where we also have adequate knowledge about the curricula used. Whether the method can be extended to other areas of Sápmi remains to be seen.

The structure of the paper is as follows. We start by presenting some figures regarding the present-day number of speakers/users of the Sámi languages in general. We then hone in on the three varieties currently in use in Norway and their status in the Norwegian educational system. On the basis of the numbers presented, a set of prognoses is put forth and discussed before aspects of a broader model to assess the future of the Sámi languages in Norway are taken into consideration.

[1] The spelling 'Saami' is frequently encountered in English texts and used for instance by both Ethnologue and the Endangered Languages Project. In the present text we use 'Sámi' with accent aigu on the root vowel which better reflects the spelling used in several of the Sámi languages themselves.

 Nordlyd 46.1: 259–272, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60*, edited by Lene Antonsen, Sjur Nørstebø Moshagen and Øystein A. Vangsnes. Published at UiT The Arctic University of Norway. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6427

2. The Sámi languages: facts and figures

2.1 Number of users

The Sámi languages constitute a separate group within the Uralic language family. The language group makes up a dialect continuum in which nine different main varieties are recognized as living languages today. Table 1 lists them from northeast to southwest with number of speakers as currently (April, 2022) given by *Ethnologue* and the *Endangered Languages Project* (ELP), respectively.

Language	# speakers Ethnologue	# speakers ELP
Ter Sámi (RU)	2	30
Kildin Sámi (RU)	600	~300
Skolt Sámi (RU, FI)	320	~300
Inari Sámi (FI)	300	~300
North Sámi (FI, NO, SE)	25,700	16,500
Lule Sámi (NO, SE)	2,000	1-2,000
Pite Sámi (SE)	20	~42
Ume Sámi (SE)	20	<20
South Sámi (NO, SE)	600	600

Table 1: Number of speakers of Sámi languages according to Ethnologue and the Endangered Languages Project

There are a few obvious discrepancies between the figures in the two databases, which presumably is due to the fact they may use slightly different sources which to varying degree may be updated. The definition of 'speaker of language x' may also vary considerably across sources and researchers, from counting just native, first language users to also including individuals with only partial knowledge of the language. Official registers of language users do not exist in all of the four countries, and notably not in Norway which hosts the highest number of ethnic Sámi. Furthermore, we also see that the ELP figures are somewhat approximate. In fact, the ELP provides alternative sources for their figures while at the same time assessing the certainty of each of them, and the ones given in Table 1 are those assessed to be most certain.[2, 3]

What is nevertheless unquestionable is the following: North Sámi has far more speakers/users than all of the other varieties combined with about 85–90% of the total Sámi speaker/user popularion. Whereas North Sámi is considered *vulnerable* by the ELP, all the other varieties are considered *endangered* to varying degrees. As a whole the Sámi languages are presumably the most threatened historical minority language group in all of Europe with its total number of speakers at best being about 30,000, possibly just about 20,000.

Crucially, due to a long period of explicit assimilation policies in all four nation states that divide up Sápmi, there has been a breach in inter-generational transmission of the language in many local communities (see e.g. Minde 2003, Trosterud 2008, Albury 2016 and references cited there). A consequence of this is that there are currently many people with a Sámi ethnic identity and/or ancestry who have little or no command of one of the Sámi languages. Pietikäinen, Huss, Laihiala-Kankainen, Aikio-Puoskari, & Lane (2010:4) estimate the total Sámi population to be between 140, 000 and 200,000. In other words, only some 10–20% of all ethnic Sámi are speakers of a Sámi language. Rasmussen (2017:43) estimates the number of ethnic Sámi in Norway to be at least 100,000, which, as we will see below, gives a similar ratio of language users to ethnic members.

[2] The source for the estimated 16,500 users of North Sámi is considered 100% certain by the ELP. However, they also cite another source with the estimate 25,700 as a 100% certain, but this source (Kejonen 2020) in turn cites the figure given in Ethnologue as but one of several estimates provided in the literature.

[3] A reviewer points out that the estimated 300 users of Inari Sámi most likely refers to the pre-revitalization stage and that the current number can be taken to be about 450 users of this variety of Sámi, see Olthuis et al. 2021: 178).

2.2 The Sámi languages in Norway

In June 1990, Norway was the first state to ratify the ILO 169 convention on indigenous and tribal nations, and Norway recognizes the Sámi as an indigenous historical minority of the country. Furthermore, by §108 of the Norwegian constitution, Norwegian authorities are obliged to facilitate the development of Sámi language, culture, and way of life. Additionally, North, Lule and South Sámi are also recognized as indigenous languages of Norway by the Norwegian State under the European Charter for Regional or Minority Languages.

Ethnologue estimates the number of users in Norway to be 20,000 for North Sámi, 1,000 for Lule Sámi, and 300 for South Sámi. The figure for North Sámi is most likely too high and seems to be a confusion of users in Norway and in all of Sápmi as 20,000 is given by several sources as the total number of North Sámi users.[4] The lowest estimate for North Sámi in the ELP is based on Salminen (2007:262) who says that "[i]n Norway, the number of speakers is above 10,000, in Sweden perhaps 5,000, and in Finland approximately 1,500."

The figure for Lule Sámi in Ethnologue is probably also too high. Salminen (op. cit.: 257) says about Lule Sámi that "[t]he number of speakers lies somewhere between 1,000 and 2,000" in Norway and Sweden, but Morén-Duolljá (2010: 58), based on sources in the Lule Sámi communities, estimates the number to be just around 650 altogether in the two countries. In personal communication, Morén-Duolljá says that a good 400 appears to be a fairly good estimate of the number of active and proficient Lule Sámi speakers in Norway.

The Ethnologue figure for South Sámi appears more accurate. Citing several recent sources NOU (2016: 298) states that in Norway at least 270 individuals speak the language well and at least 340 individuals understand the language well.

For the purpose of the present paper, we will assume that in Norway there are approximately 15,000, 400 and 300 active users of North, Lule, and South Sámi, respectively.[5] Given the estimate in Rasmussen (2017:43) that there are about 100,000 ethnic Sámi in Norway (cf. above) only around 15% of them are active users of a Sámi language.

To put these figures in perspective, the total Norwegian population is currently 5.39 million (2021 figure). The regions in Norway that are part of the traditional homelands of the Sámi (from north to south: the counties Troms og Finnmark, Nordland, and Trøndelag plus the area Nord-Østerdalen in Innlandet county) have about 958,000 inhabitants. In other words, the speakers of the Sámi languages make up a small minority, and Norwegian is the dominant language throughout the region.

A few local communities in the northernmost county Troms og Finnmark have a majority of Sámi language users (Guovdageaidnu/Kautokeino, Kárášjohka/Karasjok) and altogether 13 municipalities throughout the area (4 for South, 1 for Lule, and 8 for North Sámi), as well as the county municipalities Troms og Finnmark, Nordland, and Trøndelag, are part of the Sámi Language Administrative Area (SLAA) ('Forvaltningsområdet for samisk språk', in Norwegian). This is a governmental installment, introduced as part of the Sámi Act of 1987, which grants Sámi the same judicial status as Norwegian in the area and which commits the local/regional authorities to facilitate the use and development of Sámi.

In contrast to the highly uncertain numbers of speakers/users of Sámi overall, there exist very exact figures for the number of pupils that receive instruction in the Sámi languages in Norway, and these figures are publicly available from the online database *Grunnskolens informasjonssystem* (GSI) run by the Norwegian Directorate for Education and Training. We will shortly return to these figures.

[4] Ethnologue cites "Laakso et al. (2013)" but provides no list of references: the source probably equals Laakso et al. (2016) in which the following statement can be found (p. 137): "The number of North Sámi speakers is estimated to be around 20,000 (no official statistics exist)". Although their discussion centers around North Sámi in Norway, the statement cannot *per se* be taken to be about North Sámi in Norway only.

[5] A thorough discussion of various sources for the number of Sámi language users in Norway is found in Todal (2013). The discussion does however not end in a firm conclusion regarding the numbers.

2.3 Sámi language curricula and number of pupils in Norway

The Norwegian school system is highly centralized with nationally defined curricula. Municipalities are in charge of organizing grades 1–10 (ages 6–16), which constitute the obligatory school years. The municipalities within the SLAA are obliged to follow a variant of the national curricula which to a greater extent emphasizes Sámi matters. Furthermore, the municipalities may locally enforce Sámi language instruction as an obligatory subject in its schools, but only four of them do so (Rasmussen, 2015:18[6]). At the same time, all pupils in the area, regardless of ethnicity, have the right to learn Sámi. Outside the SLAA, only pupils with a Sámi ethnic background have the explicit right to receive instruction in Sámi language as a school subject.

As for the curricula for the Sámi language subject in grades 1–10, there are three variants, which are the same across the three linguistic varieties. The curriculum "Sámi as first language", generally referred to as 'Sámi 1', entails that the main literacy training happens through Sámi. The pupils who follow Sámi 1 receive five hours of instruction per week in the subject throughout the school years. Most of the North Sámi 1 pupils are enrolled in programs where (North) Sámi is the medium of instruction across all or most subjects, either at schools where Sámi is the main language or in Sámi medium classes/groups at Norwegian medium schools. In the school year 2020/2021 this applied to 92.7% of the North Sámi 1 pupils. For Lule and South Sámi there are no Sámi medium instruction programs.

The curriculum "Sámi as a second language" comes in two varieties, Sámi 2 and Sámi 3. Sámi 2 is a curriculum for pupils who have some knowledge of the language but who otherwise receive their main literacy training through the subject Norwegian and who have Norwegian as the medium of instruction in other subjects. They receive 3–4 hours of instruction in Sámi per week, varying somewhat with grade level. Sámi 3 is a curriculum that requires no prior knowledge of Sámi, and which is offered two hours per week throughout grades 1–10. Also for the Sámi 3 pupils Norwegian is otherwise the medium of instruction across subjects.[7]

It is worth pointing out there does not exist any language immersion program for Sámi in the Norwegian school system. The system caters for language maintenance in so far that the majority of Sámi first language children are enrolled in a program with Sámi as the medium of instruction (SMI), but there is no systematic way of including children with low competence in Sámi into such a program, for instance in the way it has been done for Inari Sámi in Finland (see Olthuis, Kivelä and Skutnabb-Kangas 2013). What we rather see is that there is a net loss of pupils from all Sámi curricula over the school years (see Vangsnes 2021), a fact that we will return to below in section 4.

Although it is tempting to link Sámi 1, Sámi 2, and Sámi 3 to the quite widely used terms 'first', 'second', and 'third' (or alternatively 'foreign') language, it is important to stress that they first and foremost signify the curricula and the extent of the formalized instruction: there may in fact be, and presumably are, pupils with Sámi as a first language who nevertheless for various reasons follow either Sámi 2 and Sámi 3.

	Sámi 1	Sámi 2	Sámi 3	Total
North Sámi	943	728	637	2,308
Lule Sámi	33	56	24	113
South Sámi	35	50	16	101
Total	1,011	834	677	2,522

Table 2: Number of pupils following different curricula for the three Sámi languages in use in Norway in the school year 2020/2021

[6] When Rasmussen (2015) published his paper, the following three municipalities had Sámi as an obligatory school subject: Guovdageaidnu/Kautokeino, Kárášjóhka/Karasjok, and Unjárga/Nesseby. In 2021 Deatnu/Tana municipality also decided to install Sámi as an obligatory subject in grades 1–4, see https://www.tana.kommune.no/cppage.6413478-161070.html.

[7] For further details, see https://www.udir.no/laring-og-trivsel/lareplanverket/kunnskapsloftet-samisk/.

Table 2 presents the number of pupils who followed the different curricula in the school year 2020/2021. The table quite clearly shows that North Sámi is by far the biggest of the Sámi languages also in the context of education in Norway.

The total number of pupils has not changed much over the last decade. In 2009 it was 2,336 decreasing to 2,116 in 2014 before increasing again. Figure 1 shows this development graphically for all three language varieties. Figures 2–4 furthermore illustrate the development for North Sámi, Lule Sámi, and South Sámi, respectively, each figure distinguishing between the three kinds of curricula. What is particularly noticeable for North Sámi is that there is a growing number of pupils following Sámi 2 at the expense of both Sámi 1 and Sámi 3 although the largest group is still Sámi 1. For both Lule and South Sámi the majority of pupils follow the Sámi 2 curriculum.

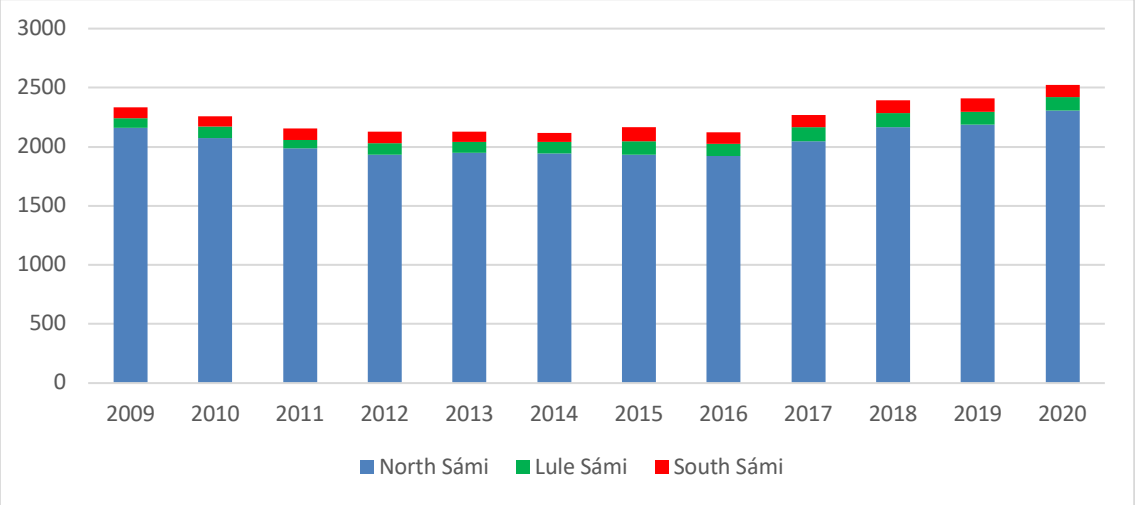


Figure 1: The total number of pupils receiving instruction in Sámi grades 1-10 from 2009 to 2020

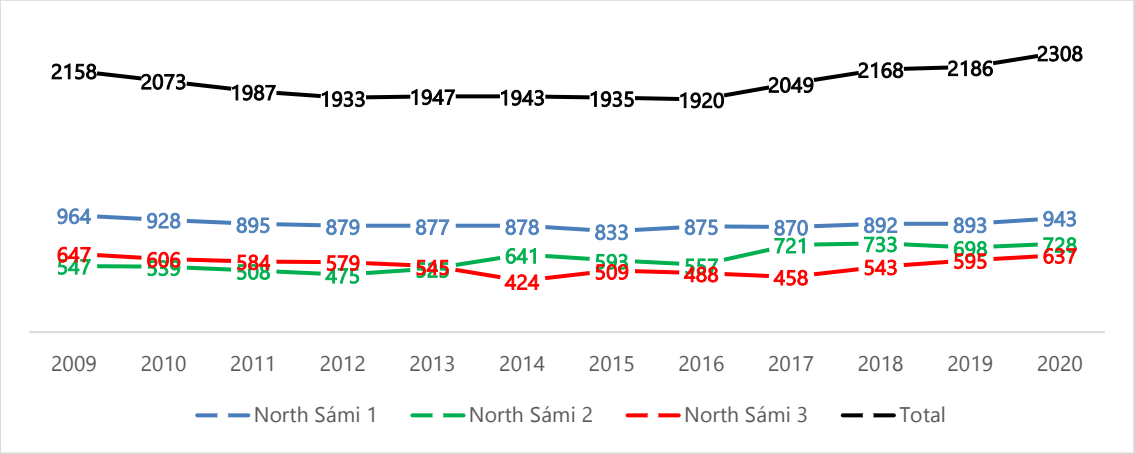


Fig. 2: Number of pupils in grades 1-10 with instruction in North Sámi in Norwegian schools between 2009 and 2020

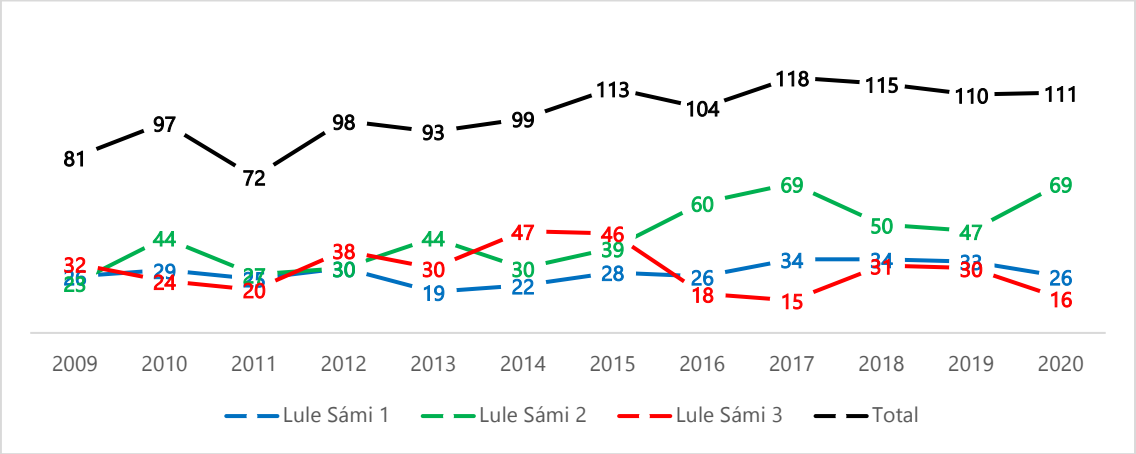


Fig. 3: Number of pupils in grades 1-10 with instruction in Lule Sámi in Norwegian schools between 2009 and 2018

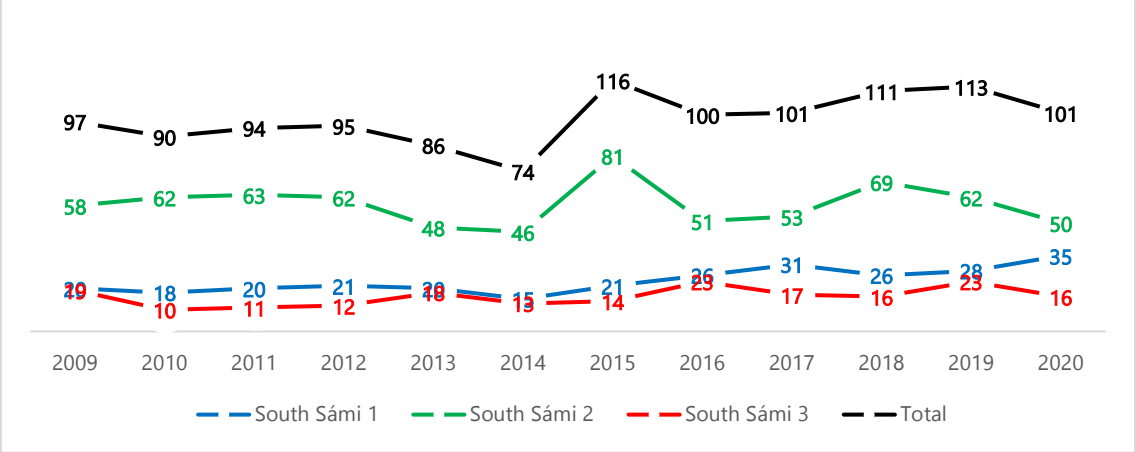


Fig. 4: Number of pupils in grades 1-10 with instruction in South Sámi in Norwegian schools between 2009 and 2018

In 2020/2021, the total number of pupils in grades 1–10 in Norwegian schools was 635,497. That means that the 2,394 pupils who received some form of instruction in one of the three "Norwegian" Sámi languages made up 0.38% of the total pupil population. The 952 Sámi 1 pupils made up 0.15% of all Norwegian pupils. This is worth keeping in mind as a token of how minoritized the Sámi languages are in the Norwegian society.

The school data just presented will form the basis for the prognoses to be developed in the next section. As mentioned in the introduction, a full model for estimating the future of a language population must also take into account recruitment of individuals who learn the language without receiving formal instruction in it in the school system. Nevertheless, basing the prognoses on the population of Sámi pupils relies on the assumption that this group will constitute the core of the future population of Sámi language users. The assumption may be scrutinized, and although ultimately an empiricial question, the following characterization of the contemporary situation for Sámi language and culture by Ulla Aikio-Puoskari (2009) may be taken in support of it:

> The present situation is characterised by an intensive struggle between a language and cultural shift on the one hand, and revitalisation and cultural survival on the other. The school and the teaching of and through the native languages are of great significance for the outcome of this competition. (Aikio-Puoskari 2009: 218)

3. Projections: The method

The total population of Norway was 5.39 million in 2021. In the school year 2020/2021, there were 635,497 pupils in grades 1–10, which means that they make up 11.8% of the total population. That gives a ratio for pupils to the rest of the population of 1:7.4.

We will now make the following assumptions: (i) the children receiving instruction in a Sámi language will be the main base of the future active speakers and bearers of the language, (ii) this number will remain at the same level as today, and (iii) the basic birth rate is more or less the same among the Sámi as in the rest of the Norwegian population. None of these assumptions are a natural given, but if we make them, we can estimate the future number of active users by the following method:

(1) *Multiply the school figures by 7.4 and add to the resulting sum the school figures, i.e. (n * 7.4) + n = x.*

In the following we will let *n* be the figures for the school year 2020/2021. An alternative using the average of the 12-year period 2009-2020 would have given slightly smaller numbers, and since the fluctuation in Table 2 may signal a rising trend, we choose the last year as the basis for the projections. Furthermore, depending on whether we think that only Sámi 1 pupils, both Sámi 1 and 2 pupils, or Sámi 1, 2 as well as Sámi 3 pupils all will become future users and bearers of the languages, we get three different prognoses which we may term 'sober', 'moderate', and 'optimistic', respectively. Only incremental combinations of Sámi 1 to 3 are worth considering, which means that we exclude the combinations Sámi 2+3 and Sámi 1+3.

Table 3 gives the projections for North Sámi, and we use the figures for the school year 2020/2021. The individual projections from Sámi 2 and Sámi 3 are given for transparency only.

	n	n * 7.4	(n * 7.4) + n
North Sámi 1	943	6,978	7,921
North Sámi 2	728	5,387	6,115
North Sámi 3	637	4,713	5,351
North Sámi 1+2	1,625	12,365	14,036
North Sámi 1+2+3	2,168	17,079	19,387

Table 3: Projections for future users of North Sámi based on pupils to rest of population (ratio 7.4)

On the estimate given in the previous section that there are about 15,000 proficient users of North Sámi in Norway today, the sober prognosis for North Sámi is that the number will be reduced by almost a half, whereas by the moderate prognosis there will be a slight decrease, whereas by the optimistic one there will be a fair increase.

Table 4 gives the projections for Lule Sámi. Again Sámi 2 and 3 are given for transparency only.

	n	n * 7.4	(n * 7.4) + n
Lule Sámi 1	33	244	277
Lule Sámi 2	56	414	470
Lule Sámi 3	24	178	202
Lule Sámi 1+2	89	659	748
Lule Sámi 1+2+3	113	836	949

Table 4: Projections for future Lule Sámi users based on pupils to rest of population (ratio 7.4)

Given the estimate that there are about 400 active users of Lule Sámi in Norway (see above), the sober prognosis indicates a reduction in number of users to less than ¾ of today's number. By the moderate prognosis, however, there will be almost a doubling of users, and by the optimistic prognosis the number will be more than doubled. In other words, the situation looks better for Lule Sámi than for North Sámi.

Table 5 gives the projections for South Sámi. The current number of active users of South Sámi in Norway is about 300 (see above). By the sober prognosis the number of future users will remain about the same as today. By the moderate prognosis, however, the number will be more than doubled, and by the optimistic prognosis the number will be almost tripled.

	n	n * 7.4	(n * 7.4) + n
South Sámi 1	35	259	294
South Sámi 2	50	370	420
South Sámi 3	16	118	134
South Sámi 1+2	85	629	714
South Sámi 1+2+3	101	747	848

Table 5: Projections for future South Sámi users based on pupils to rest of population ratio (7.4)

The projections are shown graphically for North Sámi in Figure 5 and for Lule and South Sámi in Figure 6.

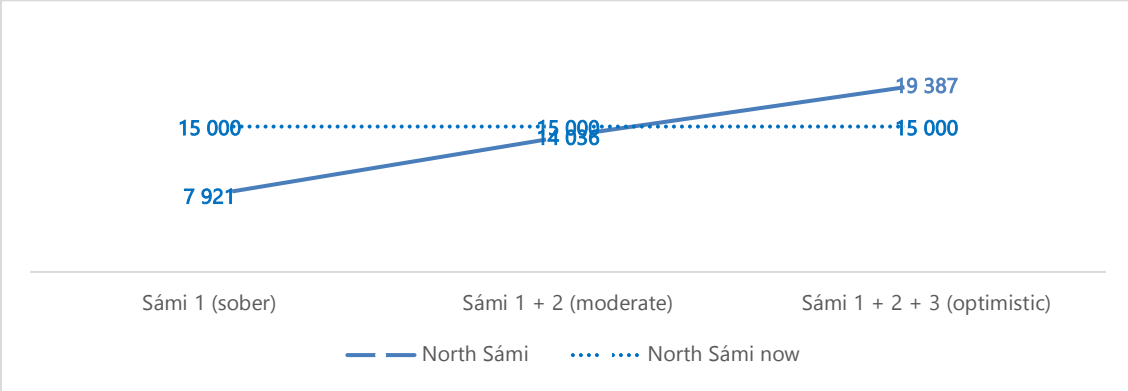


Figure 5: Future users of North Sámi in Norway projected from the current number of pupils in grades 1-10

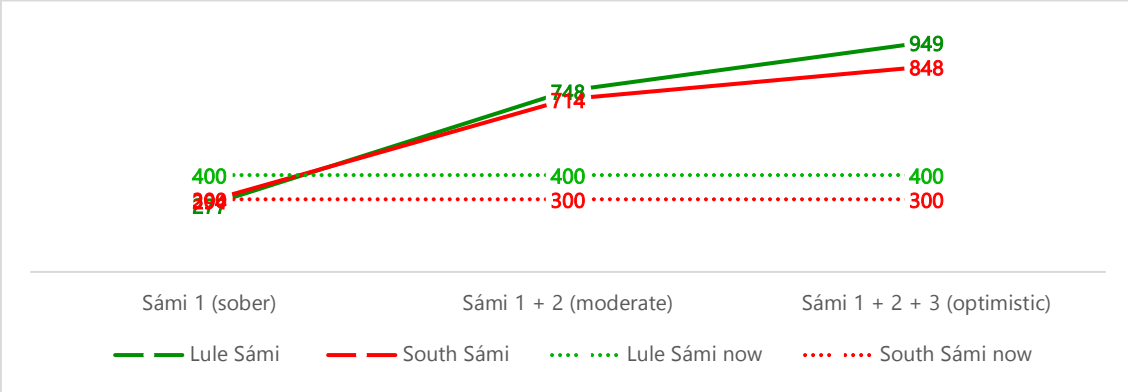


Figure 6: Future users of Lule and South Sámi in Norway projected from the current number of pupils in grades 1-10

The prognoses can be summarized as follows:

(i) *Sober prognosis*: If the Sámi 1 curriculum is the only one which by and large is successful in producing future active users and language bearers there will be a decrease in the numbers compared to today. Although the decrease is only slight for Lule Sámi and none for South Sámi, it will be quite substantial for North Sámi, and thus also for all languages viewed together.

(ii) *Moderate prognosis*: If both the Sámi 1 and Sámi 2 curriculum prove to produce future active users and language bearers the situation for North Sámi will be relatively stable, whereas for Lule and South Sámi there will be a significant increase in the numbers.

(iii) *Optimistic prognosis*: If all three curricula serve to produce future active users and language bearers, all three varieties will experience an increase in numbers, and this increase will be particularly pronounced for Lule and South Sámi with two to three times more users.

The obvious conclusion to draw from this, is that in order for the educational system to serve as a vitalizer of Sámi in Norway, more than just the Sámi 1 pupils need to develop an active competence in the languages. Ensuring a good development also for the Sámi 2 pupils will give stability when we view the three languages as a whole—the significant increase for Lule and South Sámi will be subsumed by the numerical dominance of North Sámi. If also all Sámi 3 pupils obtain an active competence in the languages there will be an overall increase.

The question is which of the three scenarios—the sober, the moderate or the optimistic prognosis—is the most realistic one, or even whether any of them are in fact adequate. A variety of factors concerning acquisition in childhood and adolescence as well as opportunities to use and develop the language in adulthood must also be considered regarding which pupils and how many of them remain active language users also later in life. Such factors must be part of a broader model for assessing the vitality and future of the languages. In the next section we will discuss the prognoses further and also sketch key properties of a broader model for projections of the future number of users.

4. Discussion: towards a broader model

We may start the discussion of a broader model by considering four issues pertaining to the school-based prognoses as such. First, the optimistic prognosis may be deemed quite unlikely. The two hours of Sámi language instruction per week provided for the Sámi 3 pupils are not sufficient to make a substantial portion of them active and proficient users of the language. Many of these pupils have little or no prior knowledge of Sámi beforehand as many of them come from homes and families where Sámi is not spoken. Unless they are exposed to Sámi in other ways and in other contexts, they are thus not likely to develop their proficiency to an advanced level. Still, one should not underestimate the role of the Sámi 3 curriculum as one providing a useful base to support exposure from extra-curricular sources and/or a base from which the individual pupil later by self-driven interest may develop a stronger language competence.

Second, it is an open question whether all Sámi 2 pupils will remain active users of Sámi later in life. Sámi 2 pupils receive 3-4 hours of Sámi instruction per week and will have some prior knowledge of the language when entering school, but Norwegian will all the same be the main medium for literacy training for them. We may assume that for many of these pupils Norwegian is likely to play an increasingly dominant role in their lives as they mature.

Third, the sober prognosis could have been made more fine-grained by distinguishing between those Sámi 1 pupils that follow Sámi medium instruction (SMI) and those who do not. Pupils in SMI will inevitably get significantly more exposure to and training in a Sámi language than the non-SMI Sámi 1 pupils. The latter group is thus in a more vulnerable situation. Fortunately, this group constitutes a minority among the first language pupils: in the school year 2020/2021, they were 69 (out of 943) North Sámi, 33 Lule Sámi and 35 South Sámi pupils, which means 13.6% of all Sámi 1 pupils.

Fourth, an additional complication represented by using the 2020/2021 sum of all pupils is that when one looks closer at individual cohorts, there is a net loss of pupils over time. As reported in Vangsnes (2021), on average 20.4% of the North Sámi 1 pupils who started first grade in the 12-year period from 2003 to 2011 left the program. For North Sámi 2, counted from a peak in pupil numbers in second grade, the loss is 33.6%, whereas the loss for North Sámi 3 is 63.1% counted from a peak in third grade. The total average loss across all curricula is 38.2% from top to bottom and 31.5% from first to tenth grade. These matters are illustrated graphically in Figure 7. The numbers for Lule and South Sámi are so small that they are not amenable for a similar statistical analysis, but a discussion can be found in Vangsnes (op. cit.).

The reasons for this loss of pupils from the different kinds of formal instruction in Sámi need to be further investigated. To what extent the "leavers" also stop using and developing their competence in Sámi, we do not know. But crucially, if they do stop, the number used in the above prognoses is too high and must be adjusted accordingly.

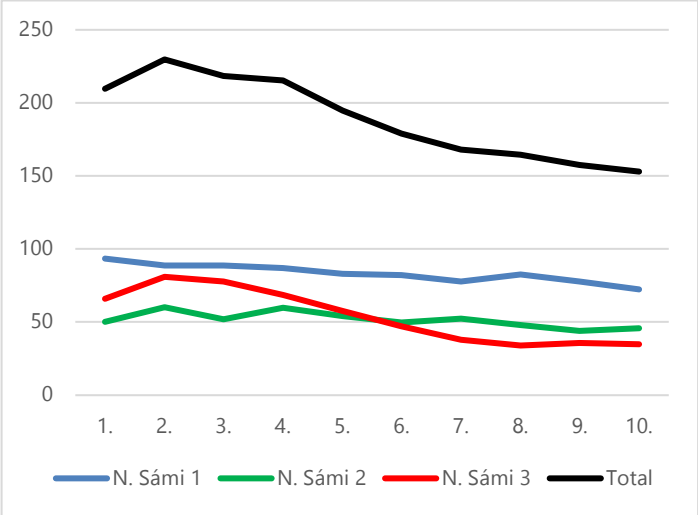


Figure 7: Average development of cohort size grades 1-10 for the North Sámi cohorts starting school in years 2003-2011. (North Sámi 3 was introduced in 2006 and grades 1-3 are therefore weighted for this curriculum.)

In addition to the four issues that have now been pointed out pertaining to the school-based prognoses *per se*, a more accurate prognosis should include groups of active language users who have acquired a functional competence in Sámi without formal instruction in the educational system. Two such categories appear especially relevant: (i) first language (L1) users who grow up using Sámi in the home environment but without receiving any formal instruction at school, (ii) second language (L2) users who have acquired the language in adulthood, i.e. the category increasingly referred to as 'New Speakers', see e.g. the papers in Soler & Darquennes (2019, and references provided there; see also Rasmus (2019), Rasmus and Lane (2021), for a qualitative study of Sámi New Speakers in two Sámi communities in Norway, and Pasanen (2019), for a discussion of Sámi New Speakers in Finland. In the advent of more thorough surveys we can only guess how big these groups of users may be in the Sámi context.

Regarding the first category, given that Norwegian Sámi L1 children are entitled to instruction in Sámi no matter where in the country they live, there are reasons to expect that rather few Sámi L1 children are not subsumed by official school statistics of the kind provided here.

On the other hand, adult L2 learners may represent a significant contribution to the population of speakers/users. In Norway, there are currently 19 Sámi language centers (see Sametinget, Undated) that organize different kinds of Sámi language courses (Nygaard et al., 2012). The courses are organized at varying intervals and they target different age groups and existing competence levels. In addition to this, both the Sámi University of Applied Sciences in Guovdageaidnu/Kautokeino and UiT The Arctic University of Norway from time to time organize courses in North Sámi as a foreign language. Some courses are completed with an exam that give university level credits and which state particular competence levels. Other courses do not.

Antonsen (2015) provides an overview of how many students have taken exams and obtained credits in the years 2009-2014 from the Sámi as foreign language courses. According to her overview (op. cit.: 76ff), which involves courses provided at eight locations, 189 students completed an introductory level course, 130 students a second level introductory course, whereas 11 students completed a one semester course at the subsequent level. However, it is not known how many of these students are adult L2 learners with no or little knowledge of Sámi from childhood. Some of the students are likely to have had some Sámi background. They may even have had formal instruction in Sámi in school and hence be included in the school statistics discussed above.

Still, given that the number of proficient Sámi users is so low in the first place, even a small number of proficient adult L2 learners each year would make a significant contribution. If we for the sake of the argument say that the 19 centers and two universities annually produce on average three proficient New

Speakers each, over a time span of 50 years that would make up (21 x 3 x 50 =) 3,150. Fifteen of the institutions focus on North Sámi, and adding 2,250 to the estimates above would make both the sober and the moderate prognosis less grim for this variety of Sámi.

The bottom line is nevertheless that more research is needed to establish a better estimate of how many adult L2 learners obtain an active competence in Sámi outside of the regular school system. At the same time, such investigations should also seek to establish in what way, and by what numbers, the course activities serve to complement Sámi language instruction in school and thereby how they support pupils in both becoming and remaining active language users also in their adult lives.

The latter point brings us to the topic of 'language vitality'. Bodies like Ethnologue, UNESCO and the Endangered Languages Project all list factors to assess language vitality which, although varying in number (from 12 to 4), largely overlap and converge. The nine factors given in UNESCO (2003) are the following.

1. Intergenerational language transmission
2. Absolute number of speakers
3. Proportion of speakers within the total population
4. Trends in existing language domains
5. Response to new domains and media
6. Materials for language education and literacy
7. Governmental and institutional language attitudes and policies including official status and use
8. Community members' attitudes toward their own language
9. Amount and quality of documentation

For each of these factors, UNESCO provides a scale from 0 to 5 to assess the vitality of a given language: the closer to 5, the more vital. In effect, a qualitative judgment of each factor can then be turned into a number which can be used to do a quantitative assessment.

It is outside the scope of the present paper to review the UNESCO assessment for the Sámi languages and/or similar protocols found in Ethnologue and the ELP. We may note that according to the UNESCO scheme North Sámi is currently classified as a *definitely endangered* language whereas both Lule and South Sámi are classified as *severely endangered* languages. What may be worth pointing out regarding this classification is that the projections provided in this paper signal more positive trends for both Lule and South Sámi than for North Sámi in the area of education.

Placing education at the center of the model, we can regard the recruitment of future active language users as a function of the number of pupils obtaining a strong productive competence in the language minus the number of such individuals leaving the language on their way into adulthood plus adult L2 learners, schematically represented as follows:

$$\{n \text{ child learners}\} - \{n \text{ leavers}\} + \{n \text{ New Speakers}\} = \{n \text{ future language users}\}$$

Vitality factors like the UNESCO ones play into this calculation either by supporting acquisition and use or by contributing to (or preventing) language shift and loss in individuals.

Returning to the Norwegian specific school system one may safely say that in order to increase the number of child learners the focus should be on the following: First, the number of Sámi 1 pupils—preferably in Sámi medium education—should be increased as this is the group which is most likely to become and remain active future users of a Sámi language. Second, the reasons for why a significant number of pupils leave Sámi instruction during the course of the obligatory school years should be given special attention. Third, the Sámi 2 and Sámi 3 curricula should be made as good as possible so as to provide good opportunities for individual pupils to strengthen their competence in Sámi outside the school system now or later in life. A fourth recommendation could be to formulate an ambition that the Sámi 3 and Sámi 2 curricula could facilitate pupils to "step up" to the more comprehensive curriculum (i.e., Sámi 3 → Sámi 2 → Sámi 1).

5 Conclusion

It should be clear from the discussion above that a number of factors need to be investigated further before more reliable projections of future numbers of Sámi language users (in Norway) can be made. We are currently not in a position to make projections like the one for example done for Welsh in Jones (2012:116).

Still, in a country like Norway it seems justifiable to assume that the number of pupils that receive literacy training in Sámi through the school system will make up a significant core of the future population of language users. Two arguments are central for underscoring this: (i) the Norwegian educational system is highly centralized and the teaching of Sámi is systematized in the form of three different curricula that target different groups of pupils, (ii) all ethnic Sámi children have a right to receive instruction in Sámi independently of where in Norway they live, and many make use of this right.

Based on this key presumption we can make three different projections—a sober, a moderate, and an optimistic prognosis—depending on whether just the first language pupils (Sámi 1) or also the second language pupils (Sámi 2) or even the foreign language pupils (Sámi 3) become future language users. By the sober prognosis North Sámi will experience a substantial decrease in numbers whereas Lule Sámi will see just a slight decrease and South Sámi hardly any. By the moderate analysis the future number of North Sámi language users will be more stable whereas the other two varieties will have a noticeable increase. By the optimistic prognosis even North Sámi will have an increase in number of users, and Lule and South Sámi will both see a very high increase.

A number of issues that may modulate the prognoses have been discussed, and although there is a high degree of uncertainty, at least one may hope that this study serves to point at some topics that need further investigation and that it highlights some aspects of how the Norwegian school system deals with instruction in the Sámi languages.

Some ethical remarks are in order at this closing point. The present author is not a member of the Sámi community and the method used in the study has not been developed in close collaboration with members of the community or with Sámi scholars, although both methods and results have been presented at several events organized by Sámi and for Sámi audiences. It should be stressed, though, that the purpose of the investigation has been to contribute positively to the efforts to strengthen the position of the Sámi languages in Norway and beyond. Hopefully, the study displays what assets, possibilities and challenges the current state-of-affairs provides, and that it thus may aid the discussion of what measures should be advanced to benefit the Sámi languages and its users in the future.

Acknowledgements

Versions of this paper have been presented at the Bilingualism Matters Research Symposium in Edinburgh in September 2019, at the Language in Context seminar at the University of Edinburgh in October 2019, at the 4th Saami Linguistics Symposium in Uppsala in November 2019, and at the FEL XXIII conference in Sydney in December 2019, and the main ideas have also been presented at a number of smaller meetings and seminars with a focus on Sámi language and culture. I am grateful to the audiences on these occasions for valuable feedback. Furthermore, I thank Lene Antonsen, Wilson McLeod, and Hanna-Máret Outakoski for more detailed discussions of matters raised in the paper, and I also thank three anonymous reviewers for their useful comments. Remaining shortcomings are my responsibility alone.

References

Aikio-Puoskari, Ulla. 2009. The ethnic revival, language and education of the sami, an indigenous people, in three nordic countries (Finland, Norway and Sweden). In *Social Justice through Multilingual Education*, edited by Tove Skutnabb-Kangas, Robert Phillipson, Ajit K. Mohanty and Minati Panda, pp. 218–239. Multilingual Matters, Bristol. https://doi.org/10.21832/9781847691910-016

Albury, Nathan J. 2016. Holding them at arm's length: A critical review of Norway's policy on Sámi language maintenance. *Journal of Home Language Research*, 1:1–16. https://doi.org/10.16993/jhlr.25

Antonsen, Lene. 2015. Språksentrenes voksenopplæring. *Sámi logut muitalit* 8: 71–84.

Jones, Hywel M. 2012. A statistical overview of the Welsh language. Welsh Language Board. Available at https://www.webarchive.org.uk/wayback/archive/20120330040554/http://www.byig-wlb.org.uk/English/publications/Publications/A%20statistical%20overview%20of%20the%20Welsh%20languagef2.pdf

Kejonen, Olle. 2020. North Saami, Čohkkiras variety (Sweden, Norway) – Language Snapshot. In *Language Documentation and Description 17*, edited by Peter K. Austin, pp. 178–185. EL Publishing, London.

Laakso, Johanna, Anneli Sarhimaa, Sia Spiliopoulou Åkermark, and Reetta Toivanen. 2016. *Towards Openly Multilingual Policies and Practices: Assessing Minority Language Maintenance Across Europe*. Multilingual Matters.

Minde, Henry. 2003. Assimilation of the Sami: implementation and consequences. *Acta Borealia* 20: 121–146. https://doi.org/10.1080/08003830310002877

Morén-Duolljá, Bruce. 2010. De samiske språkene: vakre, unike og uerstattelige. *Bårjås*, 2010: 54–65.

NOU 2016. Hjertespråket—Forslag til lovverk, tiltak og ordninger for samiske språk. Norwegian Public Report 2016:8. https://www.regjeringen.no/no/dokumenter/nou-2016-18/id2515222/

Pasanen, Annika. 2019. Becoming a New Speaker of a Saami Language Through Intensive Adult Education. In *Rejecting the Marginalized Status of Minority Languages*, edited by Ari Sherris & Susan D. Penfield, pp. 49–69. Multilingual Matters, Bristol. https://doi.org/10.21832/9781788926263-007

Pietikäinen, Sari, Leena Huss, Sirkka Laihiala-Kankainen, Ulla Aikio-Puoskari and Pia Lane. 2010. Regulating multilingualism in the North Calotte: The case of Kven, Meankieli and Sami languages. *Acta Borealia,* 27:1, 1–23. https://doi.org/10.1080/08003831.2010.486923

Nygaard, Vigdis, Áila Márge Varsi Balto, Marit Solstad and Karl Jan Solstad. 2012. Evaluering av samiske språksentre. Norut Report 2012:6. Available at https://evalueringsportalen.no/evaluering/evaluering-av-samiske-spraaksentre/Evaluering%20av%20samiske%20språksentre.pdf/@@inline

Olthuis, Marja-Liisa, Trond Trosterud, Erika Katjaana Sarivaara, Petter Morottaja, and Eljas Niskanen. 2021. Strengthening the literacy of an Indigenous language community: methodological implications of the project *Čyeti čälled anarâškielân*, 'One Hundred Writers for Aanaar Saami'. In *Indigenous research methodologies in Sámi and global contexts*, edited by Pirjo Kristiina Virtanen, Pigga Keskitalo, and Torjer Olsen, pp. 175–200. Brill. https://doi.org/10.1163/9789004463097_008

Rasmus, Sini. 2019. Sámi ođđahállit. Sosiolingvisttalaš guorahallan: go sápmelaš sámástišgoahtá rávis-olmmožin. Master thesis, Sámi University of Applied Sciences.

Rasmus, Sini, and Pia Lane. 2021. New speakers of Sámi: from insecurity to pride. *Linguistic Minorities of Europe Online.*

Rasmussen, Torkel. 2015. Samisk språk i grunnskolen og videregående opplæring. In *Sámi logut muitalit* 8. Sámi University of Applied Sciences. Available at https://samilogutmuitalit.no/sites/default/files/publications/2_sprak_i_skolen_2015_16.pdf

Rasmussen, Torkel. 2017. Eanet sápmelaččat go goassege ovdal: Leago vejolaš mihttidit ovtta giela dili. *Dutkansearvvi dieđalaš áigečála* 1/2017, pp. 39–53.

Salminen, Tapani. 2007. Europe and North Asia. In *Encyclopedia of the World's Endangered Languages*, edited by Christopher. Moseley, pp. 211–281. Routledge, London.

Sametinget. Undated. Samiske språksentre. https://sametinget.no/sprak/samiske-spraksentre/.

Soler, Josep and Jeroen Darquennes (eds.). 2019. *Language Policy and 'New Speakers'*. Thematic issue of *Language Policy*, Vol. 18:4. https://doi.org/10.1007/s10993-018-9504-4

Todal, Jon. 2013. Quantitative changes in the status of the Sámi language in Norway: A summary of existing knowledge. In *Sami Statistics Speak* 6. Available at https://samilogutmuitalit.no/sites/default/files/publications/5quantitative_changes_in_the_status_of_the_sami_language_in_norway_3.pdf

Trosterud, Trond. 2008. Language assimilation during the modernisation process: Experiences from Norway and north-west Russia. *Acta Borealia,* 25:2, 93–112. https://doi.org/10.1080/08003830802496653

UNESCO. 2003. Language Vitality and Endangerment. UNESCO Ad Hoc Expert Group on Endangered Languages. Paris. Available at http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/Language_vitality_and_endangerment_EN.pdf

Vangsnes, Øystein A. 2021. Den samiske lekkasjen i grunnskulen. In *Sámi logut muitalit* 14. Available at https://samilogutmuitalit.no/sites/default/files/publications/den_samiske_lekkasjen_i_grunnskulen_-_2021_k2.pdf

Conrad Svendsens beskrivelse av norsk tegnspråk

Arnfinn Muruvik Vonen
OsloMet – storbyuniversitetet

Abstract

This text presents an introductory investigation of Conrad Svendsen's analysis of Norwegian Sign Language, as it appears in a set of handwritten notes that have been preserved after him, supposedly from around 1910. The notes are significant for the history of the field since little material has been preserved about Norwegian Sign Language before the late 20th century. Svendsen was an important personality, at first in Christiania's [Oslo's] community for the education of deaf children, then in the Church's services for the deaf community, and finally in the establishment of a "home for the deaf". Since Svendsen's own sign language political views have paradoxical features, it is interesting to find out to what extent he really had an understanding of what sign language is. The introductory investigation concludes that Svendsen had a good understanding of many aspects of Norwegian Sign Language and was able to articulate well his pedagogical presentation of the material. However, he seems to assume that the sign language is less conventionalized than we have a reason today to believe that it may have been in his day. And to be able to assess how autonomous his understanding was, a closer investigation is needed of how he may have been influenced by earlier authors, not least authors writing in German.

Keywords: Norwegian Sign Language, history of sign language linguistics, Conrad Svendsen, language documentation

1. Innledning

De fleste språk i verden har et beskjedent antall språkbrukere, og de har i liten grad blitt dokumentert gjennom forskning. Norsk tegnspråk har lenge vært et slikt språk. Språket ble sannsynligvis til i forbindelse med opprettelsen av et undervisningstilbud for døve barn i Trondheim på begynnelsen av 1800-tallet og har en ubrutt historie fram til i dag. Men dokumentasjonen av språket har vært mangelfull helt fram til nå.

Denne teksten presenterer en innledende undersøkelse av Conrad Svendsens analyse av norsk tegnspråk. Materialet er et sett med håndskrevne forelesningsnotater. Svendsens analyse blir gjennomgått med to overordnede spørsmål for øyet: Hvordan ble norsk tegnspråk framstilt i faglige sammenhenger tidlig på 1900-tallet? Og hva slags holdning hadde Svendsen til tegnspråk?

2. Conrad Svendsen og norsk tegnspråk

Conrad Svendsen (1862–1943) er en markant og viktig skikkelse i døves historie i Norge. Han var selv hørende, født i Bergen og oppvokst der og i Ålesund, og han kom som ung mann til Kristiania for å studere teologi. Mens han ennå studerte, arbeidet han som lærer ved de to døveskolene som eksisterte i Kristiania på den tiden, først ved Christiania Døvstumme-Institut fra 1883 til 1886 og deretter ved Fru Rosings Taleskole for Døvstumme fra 1886. På slutten av 1880-tallet fikk han anledning til å foreta en større studiereise til ulike døveskoler på kontinentet. Etter ønske fra De Døves Forening (i dag Oslo Døveforening) ble han i 1893 ordinert til Norges første prest for døve, selv om han ennå ikke hadde fullført teologistudiet. I 1895 ble ansvarsområdet hans utvidet fra Kristiania til hele Norge. I 1898 grunnla han Hjemmet for Døve, som i dag lever videre under navnet Stiftelsen Signo og «gir tegnspråklige tilbud til døve, hørselshemmede og døvblinde på vegne av det offentlige» (Signo, u.å.). Stiftelsen består av en rekke virksomheter rundt om i Norge, hvorav den eldste, på Nordstrand i Oslo, bærer navnet Signo Conrad Svendsen senter (CSS). Det skulle framgå av det ovenstående (hentet fra Sander 2021, Lid 2018 og Vonen og Schröder 2018) at Svendsen var en sentral ressursperson og institusjonsbygger i landets miljø av tegn-

 Nordlyd 46.1: 273–284, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigert av Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publisert ved UiT Norges arktiske universitet. http://septentrio.uit.no/index.php/nordlyd https://10.7557/12.6392

språkbrukere. Ut fra hans posisjon er det interessant at hans forhold til tegnspråk, som vi skal se, framstår som tvetydig.

Her trengs litt bakgrunn for lesere uten kjennskap til norske døves og norsk tegnspråks historie: Norsk tegnspråk er et av Norges eldste og største minoritetsspråk, med en historie som kan følges tilbake til Trondheim rundt 1815, men samfunnets anerkjennelse av språket lot vente lenge på seg. Utover på 1800-tallet og gjennom det meste av 1900-tallet var, i Norge som i mange andre land, det såkalte oralistiske synet dominerende i døveskolene. Dette gikk ut på at undervisningsspråket for døve burde være (norsk) talespråk. Først fra 1980-tallet, støttet av moderne lingvistisk forskning, ble språkpolitikken endret og tegnspråk fikk igjen innpass i klasserommene. . Lesere som ønsker å lære mer om norsk tegnspråk, henvises til Halvorsen (2020) og Vonen (2020).

Svendsen var i sine unge år lærer i skoler som var tuftet på oralismens prinsipper om å unngå bruk av tegnspråk i opplæringen. Da han valgte å gå over til Rosings taleskole i 1886, var det endatil fordi han mente at undervisningen der «i sterkere grad var basert på *talemetoden*, som han på den tiden hadde mest tro på» (Sander 1994, s. 14). Det er en slående forskjell mellom dette og hans senere virksomhet i både Døvekirken (fra 1894) og Hjemmet for Døve (fra 1898). Disse ble hans viktigste arenaer, og de framstår på Svendsens tid som vennlig innstilt overfor tegnspråklig kommunikasjon, iallfall i sammenligning med skolene som hadde vært hans tidligere arbeidsplasser: «I kontrast til sin oralistiske bakgrunn står Conrad Svendsen som en bauta i en tid da døveundervisningen ville gjøre døve orale uten tegnspråk.» (Vonen og Schröder 2018, s. 149) Og allerede mens han arbeidet ved døveskolene, fikk han privatundervisning i tegnspråk av sin døve lærerkollega ved Christiania Døvstumme-Institut, Erik Strangestad (Vonen og Schröder 2018, s. 146).

På denne bakgrunn er det interessant å sette seg inn i hva Svendsen selv tenkte om tegnspråk. I tillegg kommer at det rett og slett finnes svært få forsøk på språkvitenskapelige analyser av norsk tegnspråk før 1980-tallet. Et hederlig og lærerikt unntak er Sigvald Skavlans grundige gjennomgang av språket fra 1875, som er gjenopptrykt og gjort tilgjengelig på Internett på 2000-tallet (Skavlan 2002 [1875]).

I arkivet til Oslo Døveforening har det ligget en bunke med håndskrevne papirer etter Svendsen. Papirene ble forært til Svendsens oldebarn Marit Vogt-Svendsen, da hun som den første i Norge disputerte til doktorgraden på en avhandling om norsk tegnspråk i 1991 (Vogt-Svendsen 1990). Vogt-Svendsen har vært så vennlig å la meg få tilgang til papirene.

Selv om notatene, så vidt jeg kjenner til, ikke tidligere er analysert grundig, er de omtalt flere steder. Schröder (1993, s. 2) skriver: «Vi har etter Svendsen funnet notater til forelesninger for døvelærerne om tegnspråk, der han bl.a. hevdet at tegnspråk virkelig har syntaks (setnings-oppbygning). Notatene tyder på at han forsto hva tegnspråk var for noe.» I Vonen og Schröder (2018, s. 148) anslås notatene til å være fra cirka 1910.

3. Materiale og metode

Materialet består av en perm med plastlommer, og i plastlommene er det til sammen 26 beskrevne linjeark, de fleste på ca. 17 x 12 cm. De fleste arkene er beskrevet bare på én side. Noen av tekstene er skrevet med blyant, andre med penn. Håndskriften er vanskelig å lese for et øye fra 2000-tallet. Noen av de blyant-skrevne tekstene bærer preg av å være kladder, med utstrykninger og tillegg mellom linjene. Disse tekstene er også jevnt over vanskeligere å tyde enn de tekstene som er skrevet med penn og virker ferdigere. Det er imidlertid ikke et skarpt skille mellom «kladder» og «ferdige tekster».

Rettskrivningen er nokså konsekvent, med små forbokstaver og uten dobbeltkonsonant sist i ord. Ordene er stort sett stavet på dansk måte, men en rekke norske trekk finnes, f.eks. *som* og ikke *der* som relativsetningsinnleder, *bli/blir* i stedet for *blive/bliver*, som oftest *skj-* og *gj-* i stedet for *sk-* og *g-* i ord som *forskjellig* og *gjøre*, men f.eks. *ske* (både verbet og substantivet). I alle de tekstene som blir gjennomgått nedenfor, bruker Svendsen *aa* og ikke *å*.

Flere av tekstene går over flere ark, og da har Svendsen sidenummerert dem øverst til høyre på arket. Flere av dem er også utstyrt med overskrifter, som også av og til er gjentatt (i forkortet form) ved siden av sidetallet. Riktignok er det tre av tekstene som har samme overskrift («Tegnsproget»). Med sidenummerer-

ingen og overskriftene er det relativt lett å sortere arkene i følgende tekster, her ordnet etter den rekkefølgen de står i i permen:

1. *Tegnsproget* (penn; 4 sider, hvorav den siste er bare på fire linjer)
2. *Tegnenes rækkefølge (syntax)* (penn; 3 sider)
3. *Tegnsproget* (penn; 1 side)
4. *Symbolske tegn* (blyant; 1 side)
5. *Tegnsproget* (penn; 4 sider, hvorav den siste er bare på fem linjer)
6. *Begrebskategorier (Ordklasser)* (blyant; 1 side)
7. *De efterlignende gebærder* (blyant; 1 side)
8. *Begrebsklasser* (penn; 4 sider + en setning på baksiden av første ark)
9. *De døves tegnsprog* (penn, men det meste av siste side med blyant; 3 sider)
10. *Kruse: Der Taubstumme* (penn; 2 sider, en på hver side av arket) (bruker *å*)
11. (Uten overskrift, stikkord om døve) (blyant; 1 side)
12. *Døvst. i alm.* (penn; 1 side)
13. (Uten overskrift, stikkord om døvehistorie) (blyant; 1 side)

De 13 tekstene kan sorteres i to hovedkategorier:

- Tegnspråk (tekst 1–9)
- Døves særtrekk, opplæring og historie (tekst 10–13)

For å kunne gjøre tekstene tilgjengelige for analyse måtte jeg først transkribere dem. Dette var et tidkrevende og vanskelig arbeid, og det er fortsatt detaljer i transkripsjonene som ikke er løst. Både originaltekstene og transkripsjonene vil snart gjøres offentlig tilgjengelig på Internett (Vonen, under arbeid) og vil være nyttige å oppsøke for den som ønsker mer nøyaktig tilgang til hva Svendsen faktisk skrev.

Sitatene er så nøyaktige gjengivelser som mulig av Svendsens tekster, med ett hovedunntak: Forkortelser er endret til fullstendige uttrykk der jeg har vurdert dette til å øke lesbarheten, og åpenbare skrivefeil er rettet. Disse endringene er vist med hakeparenteser. For eksempel finner vi i sitatet fra tekst nr. 1 nedenfor uttrykkene «bestand[dele]», «ansigtbev[ægelser]» og «uforst[aaelige]» der originalen har henholdsvis «bestand.», «ansigtbev.» og «uforst.». Enkelte steder har jeg ikke godt nok grunnlag til å avgjøre hva en forkortelse står for, og da har jeg latt den stå som forkortelse i sitatet. Et eksempel på det er «f.g.» i sitatet fra tekst nr. 2 nedenfor. Et eksempel på retting av skrivefeil finnes i sitatet fra tekst nr. 5, der originalens form «hoveformer» er endret til «hove[d]former». I tillegg har jeg latt være å gjengi utstrøkne og på andre måter rettede elementer i Svendsens tekst. Enkelte steder har jeg ikke klart å tyde hva Svendsen har skrevet. Da har jeg markert det som mangler i sitatet, med «[...]» (et eksempel er å finne i et sitat fra tekst nr. 5 nedenfor).

Det er mange måter å gjengi tegnspråk på i skrift. I dette kapittelet har jeg fulgt det systemet for enkel glossing av tegnspråklige eksempler som er angitt i Vonen (2020, s. 10–11). Et hovedprinsipp i slik presentasjon av eksempler er at ett tegn glosses som ett norsk ord skrevet med versaler (majuskler, store bokstaver), vanligvis en omtrentlig oversettelse til norsk av tegnet. For eksempel er VANN en representasjon av et tegn som betyr «vann». Et vanlig alternativ til majuskler i glossing av tegn er kapiteler (VANN). Morfologiske varianter av tegn samt retningsangivelser i tegnrommet representeres med tegnet «+» og en kort angivelse, for eksempel «PEK+mottaker», som angir at tegnet PEK rettes mot samtalepartneren. Når det er Svendsen selv som gjengir det tegnspråklige uttrykket, siterer jeg naturligvis hans formuleringer.

Der det er grunnlag for det, har jeg kort sammenholdt Svendsens analyser og betraktninger med hvordan de tegnspråklige fenomenene han er opptatt av, omtales i vår samtids forskning. Hensikten med dette er både å hjelpe en leser som ønsker å se Svendsens arbeid i lys av dagens perspektiver, og å undersøke i hvilken grad Svendsens tekster kan leses med faglig interesse den dag i dag. I denne sammenholdelsen har jeg primært brukt Halvorsen (2020) og Vonen (2020) som kilder, da de er de to mest helhetlige nyere framstillinger av norsk tegnspråk.

I det følgende skal vi se nærmere på de tekstene som tar for seg tegnspråk. De fire tekstene som tar for seg andre temaer, blir ikke omtalt videre i denne artikkelen.

4. Gjennomgang av tekstene

Nedenfor blir tekstene presentert, og innholdet i dem referert og kommentert, i den rekkefølge de er plassert i arkivpermen fra Oslo Døveforening.

4.1 Tekst nr. 1 «Tegnsproget»

Denne teksten åpner med å erklære: «Det vi kalder tegnsprog bestaar af 2 væsentlige bestand[dele] [,] mimik (ansigtbev[ægelser]) og haand[-] el. kropbevægelser. Uden mimik er mange tegn uforst[aaelige].» Her kan vi legge merke til at denne fonetisk baserte todelingen av språket minner om dagens tegnspråkforsknings inndeling i manuelle og ikke-manuelle deler av det tegnspråklige uttrykket. Men mens Svendsen grupperer kroppsbevegelsene sammen med håndbevegelsene, blir kroppsbevegelsene i moderne tegnspråkforskning gruppert sammen med ansiktsbevegelsene i hovedkategorien ikke-manuelle (eller non-manuelle) deler (Vonen 2020, s. 51–54).

Resten av teksten går gjennom de ulike delene av «mimikken»: munn, øye, nese, og deretter kinn. Svendsen legger vekt på sammenhengen mellom de ulike ansiktsmusklene og kroppslige funksjoner, for eksempel reaksjoner på ulike smaker.

4.2 Tekst nr. 2 «Tegnenes rækkefølge (syntax)»

Svendsen starter med å polemisere mot dem som mener at tegnspråk ikke har syntaks:

> Der siges ofte, at tegnsproget ikke har nogen syntax, ingen bestemt sætningsbygning. Det har sin grund deri, at f.ex verbet ofte helt er borte. – og naar der intet verbum er, er der ingen sætning. Saal[edes] kan den døve sige f.g. mig et æble – æble for mig. – her mangler verbet, men det er blot tilsyneladende idet man overser det mimiske udtryk, som ledsager de øvrige tegn.

Her er Svendsen på linje med sin forgjenger Skavlan (2002 [1875], s. 42), som i sin behandling av syntaksen inkluderer følgende passasje: «Hvert Sprog har sin Syntax. Ogsaa Tegnsproget har sin.» Svendsen fortsetter sin framstilling med å sette fram følgende regler for leddrekkefølge:

- «den almindelige regel er at subjektet staar først.»
- «Er der til subjekt[et] knyttet en adjektivisk bestemmelse kommer denne etter subst[antivet] (mand stor).»
- «Objektet står foran det verbum, som det er afhængigt af.»

Svendsen illustrerer de tre reglene med setningen «Manden sint gutten slaar.». Her står altså subjektet først, adjektivet «sint» står etter substantivet «manden», og objektet «gutten» står foran verbet «slaar».

Vi har ennå i dag ikke solid kunnskap om strukturen i norsk tegnspråk med hensyn til de fenomenene Svendsen omtaler her, altså grunnleggende leddrekkefølge. At subjektet står først i setningen (med unntak nevnt nedenfor), antar vi gjelder som en hovedregel, også i dag. Beskrivende adjektiv vil vi i dag si kan stå foran eller etter substantivet, men vi har fortsatt ingen forskning som kan fortelle oss når vi finner den ene eller den andre av disse to rekkefølgene. Når det gjelder rekkefølgen av verb og objekt, har Erlenkamp (2011) funnet at objektet gjerne forekommer foran verbet i ytringer der det er en lokalisasjonsrelasjon mellom de to, mens det snarere følger etter verbet i ytringer der det ikke er noen slik relasjon. Vonen (2020) problematiserer denne generaliseringen ved å påpeke at et «tungt» (komplekst) verb kan følge etter objektet, også når det ikke foreligger en lokalisasjonsrelasjon mellom objektet og verbet.

Deretter skriver Svendsen at et eventuelt adverbialt uttrykk knyttet til verbet kommer etter verbet, «hvis det da ikke allerede har faaet sit udtryk i selve verbet (slaa hæftig, slaa ofte)» eller uttrykkes ved mimikk samtidig som verbet («han hilste venlig»). I Vonen (2020) omtales disse alternativene som morfologiske prosesser i verbet, henholdsvis reduplikasjon og ikke-manuell modifikasjon. Noen tegnspråkforskere har kalt sistnevnte prosesser «orale adverb», selv om de er uttrykt som ikke-manuelle prosesser som er samtidige med håndbevegelsene i et tegn, og ikke er egne tegn.

Dermed ender Svendsen opp med et setningsskjema som han formulerer slik: «S.A.O.V.A'» [subjekt/substantiv – adjektiv – objekt – verb – adverbial].

Svendsen erklærer deretter: «Denne sætningsbygning finder sted overalt, hvor et naturlig tegnsprog opstaar – og behersker den døves tankegang.» I dag virker dette som en bombastisk påstand, da vi vet at

ulike tegnspråk har ulike leddstillingsregler. I tillegg vil vi i dag være mer forsiktige med å påstå at et grammatisk fenomen «behersker» språkbrukernes tankegang.

Eksemplene Svendsen anfører for å illustrere sitt poeng, inneholder overraskende «frie» oversettelser fra norsk sett i forhold til at de bare skal vise S.A.O.V.A'-mønsteret (vi kan her også se en vilje til moralsk oppbyggelse i eksemplenes innhold):

Læreren er gaaet ud i haven.	→	Lærer have gaa.
Læreren er forstandig og flittig.	→	Lærer klog, læse, skrive, arbeide.
Regnet gjør landet frugtbart.	→	Regnet falder, planterne vokser.
Jeg maa elske og agte min lærer.	→	Jeg slaa, bedrage, ikke lærer, jeg elske og ære.

Forklaringen på de noe eiendommelige eksemplene, viser det seg, er at han har oversatt dem direkte fra tilsvarende tyskspråklige eksempler hos Wundt (1904, s. 213). Dette er et av de få stedene i notatene hvor Svendsen angir en faglig kilde i parentes: «efter Wundt».

I tillegg skriver Svendsen at objektet kan stå først i setningen, men understreker at dette ikke er «den alm[indelige] regel», men har sin grunn i at gjenstanden «gjorde et overvældende indtryk», eller «fordi et andet begreb optager tankerne – vand drikke jeg». Her kan vi regne med at det er temaledd (topikalisering) han mener. I dette tilfellet er tegnet VANN temaledd og angir hva setningen handler om, mens DRIKKE JEG er det som utsies om vannet. Leddrekkefølgen VANN DRIKKE JEG er helt i tråd med det man vil forvente også i moderne norsk tegnspråk når VANN er temaledd. Merk for øvrig at Svendsens regler ikke gjør rede for hvorfor subjektet, som i dette tilfellet er et pronomen, står til slutt i setningen. (Om etterhengt subjektspronomen i norsk tegnspråk, se del 4.6 nedenfor om tekst nr. 9.)

Svendsen oppsummerer «[d]e grundlinjer, som kan optrækkes for tegnsprogets ordstilling», slik: «tegnsproget beretter begivenheder nøiagtig i den rækkefølge, i hvilken de opleves. Det beskriver gjenstande nøiagtig i den orden, i hvilken de paatrænger sig opmærksomheden.» Når det ikke foreligger noen tidsrekkefølge i betydningen, som for eksempel i den innbyrdes plasseringen av substantiv og adjektiv, og mellom verb og objekt, blir regelen «at den forestilling gaar først, som kan tænkes alene uden den anden og at den følger efter, som i angjældende tankeforbindelse trænger den første». Dette illustrerer Svendsen med følgende eksempel: «Et stort hus. Hvad er det faste, subst[antivet] – størrelsen variabelt. huset kan tænkes i forb[indelse] med andre bestemmelser. Bygmesteren bygger huset – huset er det faste, uden hvilket bygge ikke forstaaes.» En skeptisk nåtidig leser vil her kunne innvende at det skulle være godt mulig å tenke seg en byggeprosess uten å vite hva resultatet av prosessen skal være.

4.3 Tekst nr. 3 «Tegnsproget»

Dette er en kort tekst på bare én side. Øverst til høyre på siden har Svendsen skrevet, på tre linjer: «nydannelser / forvandlinger / forældet tegnsprog». Notatet handler om endringer i språket og begynner slik:

> Tegnsproget er ikke noget afsluttet sprog. Det opstaar paanyt hvergang der er en døv, som kommer i forb[indelse] med en anden døv, eller hverg[ang] de hørende prøver at sætte sig i forb[indelse] med en døv. – Og tiltrods for at tegnsproget saaledes dannes paanyt paa forskjellige steder er tegnene som benyttes af disse forskj[ellige] saa lige, at de forskj[ellige] personer med lethed kan gjøre seg forstaaet af hverandre.

Svendsens betraktninger her gjenspeiler det generelle paradoks at tegnspråk – i likhet med talespråk – både varierer og gjør kommunikasjon mulig. Individuell variasjon kan være noe mer påfallende i et tegnspråk enn i et talespråk, og dette vil vi i dag se i lys av forskjellene mellom de to språktypene med hensyn til overføring mellom generasjonene: De fleste tegnspråklige barn må ut av familien for å møte og sosialiseres med andre tegnspråklige. Svendsen går imidlertid lenger i sine refleksjoner på dette punktet og fortsetter ved å sette fram følgende påstand, som i våre dager vil være mer kontroversiell:

> Denne forstaaelse indskrænker sig dog væsentlig til samtale om konkrete ting – saadanne som der kan peges paa, og saadanne, hvis form eller bevægelse let kan efterlignes, eller følelser, som kan afbildes ved ansigts-udtrykket, f.ex. mand, hus, blomst, gaa, staa, ligge, stille, pen o.s.v. Derimod er der en hel del betegnelser, som er forskjellige, saadanne som omfatter personer – navn – og saadanne tegn, som kun betegner en enkelt egenskab ved en gjenstand.

Svendsen antyder altså at tegnspråkbrukere ikke nødvendigvis forstår hverandre på annet enn et konkret nivå. Dette vil vi i dag bestride, i alle fall når det gjelder veletablerte tegnspråksamfunn som det norske, der det naturligvis til stadighet foregår både akademiske og andre samtaler om et bredt spekter av abstrakte temaer.

Bestrides bør nok også følgende oppsummerende setning fra Svendsens side, i alle fall om han her har et internasjonalt perspektiv og ikke kun ser på variasjon innenfor det vi i dag kaller norsk tegnspråk: «De forskjellige tegnsprog kan betragtes som dialekter.»

4.4 Tekst nr. 4 «Symbolske tegn», 5 «Tegnsproget» og 7 «De efterlignende gebærder»

Disse tekstene utgjør en faglig helhet, der Svendsen presenterer hovedkategorier i tegnforrådet i norsk tegnspråk.

Tekst nr. 5 begynner slik: «Man deler tegnene i to store hove[d]former [,] de henvisende og de efterlignende.» Svendsens betraktninger om hva som kjennetegner de to hovedtypene tegn, er interessante, også for en nåtidig leser. Om de henvisende tegnene sier han at de kommer fra barnets gripebevegelse etter det det ikke kan nå: «efter forgjæves forsøg paa at gribe – forstaar omgivelserne, hvad det rækker sig efter. – Og barnet lærer at pege – snart paa det, som vækker nysgjerrigheden – snart for at henlede omg[ivelsernes] opmærksomhed paa gjenstanden.» De etterlignende tegnene derimot «opstaar paa et senere stadium, og det sker ved mimik som formidlende led – ansigtsudtryk fremkalder ansigtsudtryk hos barnet – sorg, glede». Videre mener han at de henvisende tegnene, «som altsaa opstaar som et stemningsudtryk», beholder sin betydning uforandret «naar de gaar over til et tegnsprog», mens dette ikke er tilfelle med de etterlignende: «De beholder ikke altid den betydning, som de havde ved sin oprindelige fremkomst som direkte følelsesudtryk». Av denne grunn, sier han, deler man de etterlignende tegnene inn i «de afmalende = efterlignende, kopierende» og de «mitbezeichnende» [medbetegnende], der de siste «omskabes i den talendes fantasi og det, som gjør indtryk paa denne, bliver det, som gir sig udtryk gjennem tegn».

Til tross for at Svendsen altså starter notatet med å dele tegnene inn i to hovedgrupper, innfører han litt senere i notatet en tredje gruppe, de «symbolske»: «de symbolske er enten henvisende el. efterlignende tegn, som gives en ny betydning – enten f.ex. forandre rum til tid – eller konkret tegn blir udtryk for et abstrakt begreb».

Videre utover i notatet går Svendsen nærmere inn på de henvisende tegnene. Han sier blant annet:

> I sin oprindelige betydning betegner altsaa de henvisende tegn en gjenstand, som er tilstede. – Men da alle de gjenstande, som omgiver os – til andre tider eller […] kan være ude af syne, saa tvinges den døve efterhaanden til for alle gjenst[ande] at lave efterlignende tegn. Paa denne maade bliver de henvisende tegn fortrængt fra sin oprindelige stilling. Det er kun to omraader hvor de henv[isende] tegn blir beholdt – til at betegne personer – jeg, du – og til at betegne rum.

Han gir så eksempler på disse to «områdene». Eksempler på henvisende tegn som betegner personer, er «jeg», «du» og «han». Eksempler på at henvisende tegn betegner rom, er «over, under, foran, bag, høire, venstre», og dessuten blant annet legemsdeler eller disses egenskaper eller funksjoner, og tegn som betegner et sted, eventuelt også overført til å brukes om tidsrelasjoner. Disse tegnene er, sier han, «naturlige udtryk, men de faar dog delvis en overført betydning, idet det henvisende gaar over i det efterlignende. – Stor, liden.» Han går også inn på at henvisende tegn for legemsdeler kan få overførte betydninger når henvisningen ikke lenger har noe med «den talende» å gjøre, for eksempel et sansebegrep. Andre eksempler på overført betydning i henvisende tegn, sier han, er når henvisning til leppen betegner fargen rød, «eller naar cistersiensermunken naar han skal betegne vin peger paa næsen – for at gjenkalde i erindring det, som gjør næsen rød».

I tekst nr. 5 går ikke Svendsen tilsvarende inn på de etterlignende eller de symbolske tegnene, men det gjør han i henholdsvis tekst nr. 7 og tekst nr. 4. Det passer derfor å se litt på disse nå. Begge disse tekstene er korte (én side hver). Sammenlignet med tekst nr. 5 har de mer preg av å være kladd – de er skrevet med blyant og ikke penn, håndskriften er utydeligere, og det er flere overstrykninger og andre endringer i teksten. Det er nærliggende å tenke seg at Svendsen hadde en plan om å innarbeide disse tekstene i den større og mer generelle tekst nr. 5.

Tekst nr. 7 om etterlignende tegn (eller «de efterlignende gebærder», som er termen han bruker i denne teksten) begynner med å konstatere følgende: «Disse frembringes saaledes, 1) at der tegnes et billede af gjenstanden eller bev[ægelsen] i luften – eller saa 2) at der fremstilles en […] form af gjenstanden med hænderne.» Han bruker henholdsvis «den tegnende og den plastiske form (tegn)» som betegnelser for disse to kategoriene av etterlignende tegn. Han sier at den tegnende framstillingsmåten «benyttes i stor udstrækning af de døve», mens den plastiske benyttes «i de tegnsprog, som har en længre tradition», som han også omtaler som «ældre tegnsprog». Det er (igjen) uklart her hva han legger i «tegnsprog» – om det er den enkelte tegnspråkbrukers versjon av et tegnspråk eller noe annet. Han skriver også at man kan bruke et «hjælpetegn» for å vise hva et tegn skal bety, og gir blant annet følgende eksempel: «have = kreds + blomst = lugte (pegef[inger] og tommelf[inger] op til næsen)».

Helt til slutt i denne teksten innfører han termen «kjendemærketegn» og sier at der en gjenstand kan framstilles på en kortere måte, skjer dette, altså at man tar ett enkelt kjennemerke i stedet for å avbilde hele gjenstanden. Blant eksemplene her nevner han at et barn betegnes som det «som bæres paa armen».

Tekst nr. 4 er svært kort og definerer symbolske tegn som «saadanne tegn, som først vækker en saadan forestilling, som kan optages med de fire sandser, for at forbinde dermed et tankeforhold, som er forskjellig fra den sandselige, men som forbindes med den via tankeforbindelser». Som eksempel anfører han blant annet at anstrengelse uttrykkes ved å tørke svetten av pannen, og at fred og det å smi vennskap uttrykkes ved å holde hendene i hverandre. I moderne lesning ville vi kanskje brukt termen «metaforisk» for Svendsens symbolske tegn.

Oppsummeringsvis kan vi si at tekst nr. 5, supplert med tekst nr. 4 og 7, inneholder interessante perspektiver på tegnspråk, selv om en nåtidig leser vil være skeptisk til den lave graden av konvensjonalisering av språket som Svendsen ser ut til å legge til grunn. Hovedinndelingen i «henvisende» og «efterlignende» tegn kan vi kjenne igjen ikke minst i våre dagers semiotisk baserte språkforskning som tar utgangspunkt i Peirces (1955) tredeling av semiotiske ressurser (Peirces *signs*) i «peking» (Peirces *indices*, jf. Svendsens «henvisende») og «avbildning» (Peirces *icons*, jf. Svendsens «efterlignende») i tillegg til «beskrivelse» (Peirces *symbols*) (se for eksempel Bø mfl. 2018; Halvorsen 2020, s. 23–25 og 97–127; Vonen 2020, s. 39–46).

Referansen til Wundt (1900) viser at ikke alt i notatet er suget av Svendsens eget bryst, og når vi konsulterer Wundts bok, ser vi at det nok også er her Svendsen har funnet inndelingen av tegnene i henvisende, etterlignende, medbetegnende og symbolske. Merk for eksempel at Svendsen endog bruker den tyske termen «mitbezeichnend» for de medbetegnende tegnene, og også at han enkelte steder bruker termen «gebærder» i stedet for «tegn» (det tyske ordet for «tegn» i tegnspråklig sammenheng er *Gebärde*). For å kunne skjelne mellom det som Svendsen har hentet fra Wundt og andre forfattere, og det han har hentet fra sitt eget kjennskap til norsk tegnspråk, trengs en nærmere analyse av likheter og forskjeller mellom Svendsens og andres arbeider enn det vi kan overkomme i denne teksten.

4.5 Tekst nr. 6 «Begrebskategorier (Ordklasser)» og 8 «Begrebsklasser»

Tekstene nr. 6 og 8 drøfter hvordan det forholder seg med tradisjonelle ordklasser i tegnspråket. Det er tekst nr. 8 som er den mest omfattende, og vi skal derfor konsentrere oss mest om den. Det er interessant at Svendsen først har gitt den overskriften «Ordklasser», men deretter skrevet inn «Begrebs» ovenfor første del av ordet.

Som i flere av de andre tekstene begynner Svendsen tekst nr. 8 med en generell konstatering: «Tegnsproget savner særskilte kjendetegn for ordklasserne.» Nærmere bestemt, skriver han, er det noen av «lydsprogets» ordklasser som ikke eksisterer i det hele tatt i «det naturlige tegnsprog», nemlig «præpositioner, bindeord og abstrakte adverbier». Substantiver, adjektiver og verb eksisterer, men de kan ikke identifiseres med en ordklasse når de brukes alene. Igjen nevner han at ting kan bli tydeligere ved bruk av hjelpetegn, og dessuten kan ulike betydninger komme fram av «den maade, hvorpaa tegnet udføres». Han nevner som et eksempel at berøring av tennene med pekefingeren kan ha fire forskjellige betydninger, avhengig av detaljer i utførelsen: «1) tanden selv 2) hvid 3) haard 4) sten. 1) der peges blot paa tanden 2) man viser alle tænder + et lyst blik 3) man slaar gjentagne gange paa tanden 4) peger paa t[anden] og gjør bevægelse for at kaste».

Videre skriver Svendsen at det kan være vanskelig å bestemme hva et tegn betyr, da både en «virksomhed», «det som virks[omheden] frembringer, gjenstanden» og «det som foretages med gjenstanden» kan ha samme uttrykk. Blant eksemplene nevner han tegnet som kan bety både «at strø» og «det som udstrøes (i alm[indelighed] salt)». Som eksempel på hjelpetegn-fenomenet: «male kaffe + drikke = kaffe, male kaffe, drikke kaffe, brun».

Han konkluderer denne delen av framstillingen med å si at det er flere grunner til at disse uttrykkene kan ha flere betydninger: For det første kan det samme tegn i forbindelsen «bære en forskj[ellig] logisk betydning, idet det snart kan være hovedbegreb, snart blot et hjælpebegreb». For det andre kan ethvert tegn som angir en handling, «være stedfortræder for et begreb, som staar i en eller anden forbindelse med handlingen». Og for det tredje kan det ha en annen betydning «ved association». Disse resonnementene er ikke helt enkle å følge, kanskje primært fordi han ikke gir noen tydelig definisjon av «hjælpetegn».

Om tidsforhold i tegnspråk skriver Svendsen: «Saaledes kan det i alm[indelighed] ikke afgjøres, om der fortælles noget, som er skeet, som sker, eller som skal ske». Til gjengjeld, skriver han, er det andre ting som kommer fram i måten et forløp rapporteres på, noe som henger sammen med «den eiendommelighed ved tegnsproget, at det når det skal berette om en handling ikke nøier sig med at fortælle, at det er skeet – men hvorledes det er skeet – verbal og adverbial blir et tegn (gaa langsomt[)]». I dag vil mange tegnspråkforskere si det så enkelt som at verb i norsk tegnspråk ikke bøyes i tid (se Halvorsen 2020, s. 63–69, og Vonen 2020, s. 81–82 og s. 99).

Til slutt i denne teksten skriver Svendsen om hvordan man uttrykker det som i talespråket uttrykkes gjennom preposisjoner og konjunksjoner. For å uttrykke dette, sier han, «maa man benytte sig dels af direkte anskuelse, dels af omskrivninger». Når det gjelder romlige forhold, bruker man direkte henvisning, for eksempel: «der sidder en kat paa berget (hvis det kan sees, bare ved at pege paa det[)]. Hvis intet af det kan sees[,] tegn for berg, antydning af retningen, kat.» Når preposisjonen står for andre forhold enn de romlige, kan man enten «overlade det til sammenhængen at anskueliggjøre forholdet (jeg tror paa Gud[)]», eller «omskrive det, slig at det kan anskueliggjøres, f.eks. Han blev hængt paa grund av tyveri = Person, tyv, hængt. Han døde, fordi han var henfalden til drik = drikke, drikke, død».

Eksempelsetningene her framstår som gjenkjennelige i moderne norsk tegnspråk, selv om tegnspråkbrukere i dag også har preposisjoner og subjunksjoner til sin disposisjon (f.eks. årsaksmarkøren som ofte glosses som SKYLD og vanligvis kan oversettes med «fordi» eller «på grunn av»).

Tekst nr. 6 er kort (bare én side), skrevet med blyant, og har knapt noe faglig innhold som ikke også inngår i tekst nr. 8. Vi kan dermed gå ut fra at tekst nr. 6 er en kladd brukt i arbeidet med tekst nr. 8.

4.6 Tekst nr. 9 «De døves tegnsprog»

Tekst nr. 9 er den mest omfattende, mest fullstendige og mest generelle teksten i samlingen. Den er en allmenn introduksjon til hva tegnspråk er, med vekt på hva som er særegent for tegnspråk sammenlignet med talespråk (eller «lydsproget», som Svendsen kaller det). Som tilfellet er med flere av de andre tekstene, er åpningen i teksten en spissformulering. I dette tilfellet får Svendsen fram både hvorfor en lærer for døve barn bør kjenne til tegnspråk (selv når språket ikke er velkomment i klasserommet), og at han forstår forskjellen mellom tegnspråk slik døve bruker det seg imellom, og de mer oppkonstruerte tegnsystemene som i varierende grad har vært laget for bruk i skoler i ulike land opp gjennom historien:

> Det tegnsprog, som en døvelærer kommer i forbindelse med paa skolen, og som enhver lærer bør have kjendskab til, om han skal kunne holde øie med, hvad der foregaar blandt eleverne, er det saakaldte naturlige tegnsprog i modsætning til det af lærerpersonalet i tidligere tider opkonstruerte. Dette tegnsprog kaldes det naturlige fordi de døve danner det selv, i det de først søger at give udtryk for de stemninger, som viser sig inden i dem og saa senere udvikles videre, naar de prøver at afbilde gjenstande og bevægelser, som føres udenfor dem. Dette naturlige tegnsprog er ikke absolut særegent for de døve, men dets væsentlige bestan[d]dele findes igjen i de hørendes gestikulation, og særlig hos de folkeslag, som er af et livligt temp[era]ment eller som staar paa et forholdsvis lavt udviklingstrin. [Understrekning med blyant i originalen.]

Vi ser at Svendsen trekker fram parallellene mellom tegnspråk og de gestene som ledsager talespråk. Mye av innholdet i dette avsnittet er i tråd med utbredte oppfatninger blant språkforskere i dag, men han ser ut til å se bort fra, eller i alle fall nedtone, at norsk tegnspråk faktisk overføres fra generasjon til generasjon:

Barn som vokser opp med tegnspråklige foreldre, kan få overført språket på en måte som er mer eller mindre parallell med generasjonsoverføring av minoritets-talespråk, mens andre barn kan få en lignende overføring i den grad de får vokse opp i et miljø av tegnspråkbrukere, for eksempel i tegnspråklige barnehager og skoler. Dessuten ville vi i en tekst av i dag naturligvis unngått den foreldede og grunnløse hierarkiseringen av folkeslag som vi ser i den siste setningen.

Videre spekulerer Svendsen over hvordan det døve barnet utvikler sitt tegnspråk ved å kombinere sine naturlige reaksjonsuttrykk med observasjon av hvordan omgivelsene forholder seg til uttrykkene og produserer sine uttrykk.

Deretter går han inn på den delen av tegnforrådet som «er afbildning af bestemte bevægelser», og konstaterer at noen av disse tegnene alternativt kan betegne bevegelsen eller den som beveger seg: «Saaledes blir skomagerens bevægelse naar han trækker i traaden med bægge hænder betegnelse baade for, at skomageren syr og for ordet skomager.» Han går så i detalj, og «i overensstemmelse med de ældre døvst[umme]lærere» (men uten å spesifisere kilden) lister han opp en interessant liste med ti «grundlag» for dannelsen av tegn, oppsummert her:

1. Tingens bevegelse blir betegnelsen: fly → FUGL, SOMMERFUGL
2. Bruken av tingen: SKJE, SPEIL, SMØR, HEST
3. Tegnet avbilder en eiendommelig bevegelse: SKREDDER, POSTBUD, DOKTOR
4. Tegnet avbilder noe man gjør med gjenstanden: LYS (blåses ut), UR (løftes opp til øret), SAU (klippes)
5. Tegnet framstiller virkningen som gjenstanden utøver på mennesket: PEPPER, EDDIK
6. Personer får navn etter karakteristisk trekk: KRISTUS («naglegaberne i hændene»)
7. Tegnet angir hvordan tingen tilberedes: KAFFE, GRØT, STRØMPE
8. Tegnet avbilder gjenstandens form: TRE, SOL, MÅNE, BORD
9. Fargen som bi-tegn: SNØ (hvit)
10. Peke på stedet gjenstanden skal anvendes: KRONE («den del af hovedet, hvor den staar»), ØRERING («den nederste del af øret»).

Oppsummerende konstaterer han «at tegnene ikke gjengiver et helhedsbillede af gjenstandene, men at de ofte kun afbilder en enkelt egenskab eller bevægelse». Av den grunn trenger døve fra forskjellige steder som møtes, noe tid til å komme overens om hvilke tegn som skal brukes, mener han. Selv om vi ikke har detaljert kunnskap om hvordan norsk tegnspråk var på Svendsens tid, vil vi nok forvente at det var noe mer konvensjonalisert enn Svendsens framstilling antyder, ikke minst fordi mange av Svendsens eksempler (og også eksemplene i Skavlan 2002 [1875]) er gjenkjennelige som tegn som brukes den dag i dag.

Deretter understreker Svendsen forskjellene mellom tegnspråk og talespråk: «Man maa, naar man skal tilegne sig tegnsproget være opmærksom paa, at det intet har med vort sprog at gjøre og at derfor ikke et enkelt tegn dækker et i sproget forekommende ord, saaledes har ordet gaa en forskj[ellig] betegnelse efter den gjenstand, som gaar – Manden gaar, hesten gaar, skibet gaar, uhret gaar o.s.v.».

Svendsen går deretter videre og konstaterer at det er sparsomt med uttrykk for verbets tider i det naturlige tegnspråk (se omtalen av tid i norsk tegnspråk i del 4.5 ovenfor). Eksemplene han kommer med for å illustrere dette, er verdifulle også fordi de illustrerer syntaktiske mønstre som vi gjenkjenner i norsk tegnspråk av i dag:

> Vil man udtrykke sætningen: Du skal ligge imorgen. Gjør man tegnet for ligge (haanden opp til kindet og hovedet bøies til side[)]; derefter tegn for imorgen (den først kommende dag) og derefter peges der paa personen. Vil man spørge om, man skal ligge den følgende dag, gjør man tegnet for ligge og saa for imorgen og sætter op et spørgende ansigt. Læg dig udtrykkes ved tegn for ligge og du idet man sætter op en befalende mine. Fortid udtrykkes ved at gjøre tegn for begrebet og dertil knytte tegn for færdig. Jeg har læst bliver lig: læse ferdig. Jeg har læst for lenge siden – læse før, før.

Eksempelsetningene her kan transkriberes omtrent slik, med Svendsens oversettelser/forklaring i høyre kolonne:

(1)	LIGGE I-MORGEN PEK+mottaker	«Du skal ligge imorgen.»
	spm	
(2)	LIGGE I-MORGEN	«spørge om, man skal ligge den følgende dag»
	bef	
(3)	LIGGE PEK+mottaker	«Læg dig»
(4)	LESE FERDIG	«Jeg har læst»
(5)	LESE FØR+gjentakelse	«Jeg har læst for længe siden»

Her er «spm» («spørsmål») og «bef» («befaling») enkle markeringer av ikke-manuelle uttrykk for henholdsvis ja/nei-spørresetning og imperativsetning, og understrekningene markerer utstrekningen av den ikke-manuelle markeringen. I både setning (2) og setning (3) er altså markeringen fastholdt gjennom hele setningen. Ellers er det enkle transkripsjonssystemet i Vonen (2020, s. 10–11) brukt.

Flere syntaktiske fenomener i disse eksemplene kan gjenkjennes i moderne norsk tegnspråk. Et eksempel er det etterhengte subjektspronomenet i setning (1) og (3), et fenomen som ikke finnes i alle tegnspråk, og som er omtalt i Vonen 2020, s. 111), og også berørt i Halvorsen (2020, s. 72). Et annet eksempel er utelatelsen av et kontekstuelt gitt subjekt (se Vonen 2020, s. 111–112) i setning (2) (underforstått PEK+mottaker) og i setning (4) og (5) (underforstått JEG).

Deretter følger et avsnitt om pronomener, der det også er anført noen setningseksempler som kan gjenkjennes som syntaktisk idiomatiske i dagens norske tegnspråk. Her nøyer jeg meg med å transkribere dem og ikke sitere Svendsens framstilling, bortsett fra oversettelsene hans til norsk:

(6)	JEG PEK+mottaker GÅ	«Skal vi gaa»
(7)	FERDIG, GÅ	«Timen er forbi, dere kan gaa ud»
(8)	VI-TO+mottaker HJELPE	«Vi skal hjælpe hverandre»
(9)	PEK-mottaker GÅ	«Du skal gaa» [understrekning i originalen]
(10)	GÅ PEK-mottaker	«Du kan gaa»

Kontrasten i leddstilling mellom setningene (9) og (10) her finner vi også i moderne norsk tegnspråk, da etterhengt subjektspronomen bare brukes i ikke-kontrastiv funksjon (Vonen 2020, s. 111). I setning (7) virker den norske oversettelsen vel omstendelig – mer representativt ville det kanskje være å oversette setningen med «Vi er ferdige, dere kan gå.», eller – i muntlig kommunikasjon – kanskje rett og slett «Ferdig. Gå.».

Beskrivelsene av kombinasjonen av første og andre persons pronomener i setningene (6) og (8) er derimot ikke umiddelbart gjenkjennelige som moderne norsk tegnspråks første person dualis-pronomen (se Vonen 2020, s. 72). Svendsen formulerer seg dessuten ulikt i beskrivelsen av dem: Om subjektet i setning (6) skriver han bare «jeg du». Om subjektet i setning (8) skriver han: «tommel og pegefinger løftes, saa peges først paa sig selv, saa paa den anden». Det er interessant at Svendsen her ser ut til å beskrive en «L-håndform» med utstrakt tommelfinger og pekefinger, mens det vanlige tegnet VI-TO+mottaker «du og jeg» i norsk tegnspråk i dag har en «V-håndform» med utstrakt pekefinger og langfinger. (Grunnen til at tegnet som tilsvarer norsk «du» er glosset som PEK+mottaker, mens tegnet som tilsvarer norsk «jeg», er glosset som JEG, er at pronomensystemet i norsk tegnspråk ser ut til å ha et skille mellom første og ikke-første person, men i utgangspunktet ikke mellom andre og tredje person. JEG er dermed glossen for første person entalls pronomen, mens PEK er glossen for ikke-første person entalls pronomnen. Disse forholdene kommer ikke Svendsen inn på.)

Avslutningen av teksten er igjen en tankevekkende refleksjon fra Svendsens side og viser hans forståelse av at tegnspråklig kommunikasjon omfatter mer enn selve tegnene:

> Det kan i det hele ikke nok fremhæves, at man, hvis man vil forstaa tegnsprogets væsen, ikke maa gaa ud fra vort sprog, man maa meget heller gaa til det paa samme maade, som man gaar til billederne i en kinematograf, thi tegnsproget er levende billeder af personer og situationer. Tegnet blir intet, hvis det ikke udføres af en person, som gaar op i det tegnet skal fremstille. Den hele person maa spille med – ansigtudtryk, kroppens bevægelser – ellers vil det ikke gribe den døve, og lidet have at sige ham.

Mange språkforskere fra vår tid vil her kunne bemerke at mange av oss i mellomtiden har innsett at ikke bare tegnspråklig kommunikasjon kan forstås bedre om vi inntar en mer helhetlig tilnærming til tegn, ansiktsuttrykk og kroppsbevegelser, men at det samme også kan gjelde for talespråklig kommunikasjon, hvis vi bare erstatter «tegn» med «ord» (se f.eks. Bø, Ferrara og Halvorsen 2018).

5. Konklusjon: Hva kan vi lære av Svendsens notater?

Innledningsvis argumenterte jeg for at en undersøkelse av Svendsens notater vil være viktig av minst to grunner: For det første finnes det nesten ingen andre eldre norske tekster som kan fortelle oss hvordan tegnspråk ble forstått før det ble politisk anerkjent slik det har blitt i dag. For det andre er det av særskilt interesse å undersøke Svendsens notater fordi han var en så sentral person i en viktig periode i norske døves historie, og fordi hans forhold til tegnspråk ut fra hans arbeid kan synes å ha vært ambivalent: Som lærer foretrakk han Rosings skole framfor døvstummeinstituttet fordi han hadde mer tro på fru Rosings renere talemetode, men som prest sto han for en linje der tegnspråk ble ønsket velkommen.

Når det gjelder det første punktet, kan vi konstatere at Svendsen hadde en god og «moderne» forståelse av mange fenomener i norsk tegnspråk. Deler av framstillingen kan minne om Skavlans (2002 [1875]), men det er ikke snakk om noen direkte kopiering av poenger fra Skavlan. Flere av eksemplene virker dessuten genuine og styrker et innarbeidet syn i tegnspråkmiljøet i Norge som går ut på at det er en sammenhengende norsk tegnspråktradisjon fra språkets begynnelse tidlig på 1800-tallet og fram til i dag. Riktignok er deler av Svendsens framstilling muligens preget av en undervurdering av norsk tegnspråk som tradert språksystem. Uansett er det sannsynligvis riktig å si at den forståelsen som Svendsen delte med sine studenter – enten de var lærerstudenter eller ferdige lærere – var preget av oppdatert kunnskap om tegnspråk for sin tid.

Hva så med det andre punktet? Hva forteller notatene oss om Svendsens holdning til tegnspråk? Det er her fristende med en overflatisk sammenligning med Skavlan (2002 [1875]). Skavlan kombinerte sine skarpe analyser av språklige forhold i norsk tegnspråk på en paradoksal måte med en fordomsfull teoretisk overbygning der han framholdt at talespråket var tegnspråket overlegent, og at tegnspråk var et fattig språk. Hos Svendsen ser vi ikke noen like tydelig slik overbygning. Riktignok påstår han blant annet at døve fra ulike steder ikke har forutsetninger for å samtale om annet enn konkrete forhold, men selv dette er skrevet i en ramme av respekt overfor språkbrukerne og deres kommunikasjonssystem.

Av hensyn til begge de anførte punktene er det viktig å følge opp denne innledende undersøkelsen med en grundigere sammenligning av Svendsens framstilling med framstillinger som Svendsen kan ha vært påvirket av, ikke minst Wundt (1900). En slik nærmere undersøkelse kan hjelpe oss å sortere mellom det Svendsen formidlet fra sin tids nyere forskning, på den ene siden, og det han bygget på sin egen personlige erfaring med og tenkning om tegnspråk, på den andre. Ved å identifisere tydeligere hva som var Svendsens egne faglige bidrag, kan vi komme nærmere både en forståelse av kunnskapsutviklingen om norsk tegnspråk i fagmiljøene før gjennombruddet for anerkjennelsen av norsk tegnspråk på 1980- og 1990-tallet, og en forståelse av mannen som forbindes med både oralistisk undervisning og tegnspråkvennlig institusjonsutvikling.

De foreløpige funnene som jeg har lagt fram i denne artikkelen, tyder på at Conrad Svendsen var faglig oppdatert på tegnspråkfeltet. Hans oralistiske lærerperiode fant sted tidlig i livet, og vi kan tillate oss å spekulere over om han i disse årene var begeistret over den faglige framgangen som ivrige oralistiske lærere håpet å se hos sine elever hvis de ble «skånet» for tegnspråk i undervisningen, men at han senere i livet i høyere grad innså hvor viktig det tegnspråklige fellesskapet er for døve mennesker.

Litteraturliste

Bø, Vibeke, Lindsay Ferrara og Rolf Piene Halvorsen. 2018. Språkøkologi. I *Tolking – språkarbeid og profesjonsutøvelse*, redigert av Hilde Haualand, Anna-Lena Nilsson og Eli Raanes, s. 61–75. Gyldendal Akademisk, Oslo.

Erlenkamp, Sonja. 2011. Grunntegnstilling i norsk tegnspråk. *Norsk Lingvistisk Tidsskrift* 29:1, 87–116. Tilgjengelig på http://ojs.novus.no/index.php/NLT/article/view/185.

Halvorsen, Rolf Piene. 2020. *Få øye på tegn. Innføring i norsk tegnspråk.* Fagbokforlaget, Bergen.

Lid, Inger Marie. 2018. Opplæring og omsorg. I *Diakoni og velferdsstat. Utvikling av en diakonal praksis i samspill med myndigheter, sivilsamfunn og borgere*, redigert av Inger Marie Lid, s. 15–36. Gyldendal, Oslo.

Peirce, Charles Sanders. 1955. Logic as semiotic. The theory of signs. I *The philosophical writings of Peirce*, redigert av Justus Buchler, s. 98–119. Dover Publications, New York. (Relevante deler skrevet ca. 1903.)

Sander, Thorbjørn Johan. 1994. *De døves kirke hundre år. 1894–1994.*Døves menighet i Oslo, Oslo.

Sander, Thorbjørn Johan. 2021. Conrad Svendsen i *Norsk biografisk leksikon* på snl.no. Hentet 21. desember 2021 fra https://nbl.snl.no/Conrad_Svendsen.

[Schröder, Odd-Inge] OIS. 1993. Verdens første døveprest i full stilling: Conrad Svendsen, født 1862, død 1943. *Nye Journal for Døve* 3.1, 1–2. Tilgjengelig på http://www.ndhs.no/wp-content/uploads/ndhs/NJD1993-1.pdf.

Signo (u.å.). *Signo.* Hentet 29. mars 2022 fra https://www.signo.no/.

Skavlan, Sigvald. 2002. *Throndhjems Døvstumme-Institut. Program, udgivet i Anledning af Institutets 50-aarige Bestaaen.* (Statped skriftserie nr. 1.) Trondheim: Møller kompetansesenter. Tilgjengelig på http://www.acm5.com/kompendier/Throndhjems_doevstumme_institutt.pdf. (1. utgivelse 1875: Lie & Sundts Bogtrykkeri, Throndhjem.)

Vogt-Svendsen, Marit. 1990. *Interrogative strukturer i norsk tegnspråk: en analyse av nonmanuelle komponenter i 86 spørsmål.* Doktoravhandling. Trondheim: Universitetet i Trondheim.

Vonen, Arnfinn Muruvik. 2020. *Norsk tegnspråk – en grunnbok.* Cappelen Damm Akademisk, Oslo.

Vonen, Arnfinn Muruvik. Under arbeid. *Conrad Svendsen: forelesningsnotater.*

Vonen, Arnfinn Muruvik og Odd-Inge Schröder. 2018. Signo og tegnspråk i lys av norsk tegnspråkpolitisk historie. I *Diakoni og velferdsstat. Utvikling av en diakonal praksis i samspill med myndigheter, sivilsamfunn og borgere*, redigert av Inger Marie Lid, s. 144–164. Gyldendal, Oslo.

Wundt, Wilhelm. 1904. *Völkerpsychologie. Eine Untersuchung der Entwicklungsgeschichte von Sprache, Mythus und Sitte. Erster Band. Die Sprache. Zweite, umgearbeitete Auflage. Erster Teil.* Verlag von Wilhelm Engelmann, Leipzig.

Mii *eai leat gal vuollánan – Vi *ha neimen ikke gitt opp: En hybrid grammatikkontroll for å rette kongruensfeil

Linda Wiechetek[1], Flammie A Pirinen[1], Børre Gaup[1], Chiara Argese[2], Thomas Omma[1]
[1]*Divvun - UiT Norges Arktiske Universitet*
[2]*Giellatekno - UiT Norges Arktiske Universitet*

Abstract

Machine learning is the dominating paradigm in natural language processing nowadays. It requires vast amounts of manually annotated or synthetically generated text data. In the *GiellaLT* infrastructure, on the other hand, we have worked with rule-based methods, where the linguistis have full control over the development the tools. In this article we uncover the myth of machine learning being cheaper than a rule-based approach by showing how much work there is behind data generation, either via corpus annotation or creating tools that automatically mark-up the corpus. Earlier we have shown that the correction of grammatical errors, in particular compound errors, benefit from hybrid methods. Agreement errors, on the other other hand, are to a higher degree dependent on the larger grammatical context. Our experiments show that machine learning methods for this error type, even when supplemented by rule-based methods generating massive data, can not compete with the state-of-the-art rule-based approach.

Keywords: Sámi language, grammar checking, neural networks, nlp, rule-based, agreement

1. Innledning

Den digitale verdenen vi lever i krever verktøy som håndterer språk. Mens dette blir oppfattet som en selvfølge for de store språkene som engelsk, spansk og en rekke andre majoritetsspråk, er realiteten for minoritetsspråk en helt annen. De fleste minoritetsspråk mangler både tastatur for å kunne skrive språket, og ordanalyse, for ikke å snakke om stavekontroll, tekst-til-tale og maskinoversetting. Nordsamisk er et av de språkene som har verktøy for både morfologisk og syntaktisk analyse, maskinoversetting og stavekontroll, og det jobbes stadig vekk med å utvikle nye verktøy. Ett av verktøyene det er behov for er en grammatikkontroll som kan være med på å øke skriftlig språkkompetanse og dermed føre til økt bruk av samisk på nettet og i den daglige skriftlige kommunikasjonen (dvs. på sosiale medier, epost, osv.).

Nordsamisk er et finsk-ugrisk språk som snakkes i Norge, Sverige og Finland og har omtrent 25 700 talere (Simons and Fennig 2018). Språktypologisk er det et syntetisk språk, der de fleste ordklassene, f.eks. substantiv og adjektiv, bøyes etter kasus, person, tall og mer. Samisk er et minoritetsspråk som konkurrerer med majoritetsspråket i et flerspråklig samfunn og trenger derfor hjelpemidler som fremmer skriftspråket – både i opplæring og administrativ sammenheng.

I denne artikkelen drøfter vi en av de mest frekvente feiltypene i nordsamisk: kongruensfeil mellom subjekt og verbal. Deretter tar vi opp den metodiske bakgrunnen for å lage en grammatikkontroll som kan rette slike feil. I neste seksjon presenteres en maskinlæringsbasert (*NeuSam*) og en regelbasert (*GramDivvun*) modell. Disse blir diskutert og evaluerert i siste delen av artikkelen.

Den regelbaserte framgangsmåten har fordelen at man kan jobbe med veldig lite tekst (tilgangen på store mengder tekst er ofte en av utfordringene for minoritetsspråk) og ha kontroll over hva de håndskrevne reglene gjør. Dekningsgraden av ulike feiltyper begrenses til de feilene man har jobbet med. Maskinlæringsmodeller behøver mye data for å bli bra. Dette kan være en utfordring for språk som samisk som ikke har tilstrekkelig med data og samtidig en rik morfologi som fører til at de enkelte formene blir sjeldnere. Data som grammatikkontroll blir trent på må i tillegg inkludere feiloppmerking, og feiloppmerking er en tidkrevende jobb. De fleste tilnærminger velger derfor å lage et syntetisk feilkorpus nettopp pga den betydelige ressursbruken. (Miłkowski 2007, Dahlmeier et al. 2013) Samtidig kan maskinlæringsbaserte metoder ha større dekningsgrad for feil man ikke har jobbet med spesifikt. Vi har oppnådd gode resultater med maskinlæring for særskrivingsfeil, dvs. lokale grammatikkfeil (Wiechetek et al. 2021). Vi ønsker derfor å undersøke nytten

 Nordlyd 46:1 285—297, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigert av Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publisert ved UiT Norges arktiske universitet. http://septentrio.uit.no/index.php/nordlyd
https://doi.org/10.7557/12.6346

og begrensningene metoden har for andre feiltyper og muligheten for å kombinere maskinlæringsbaserte og regelbaserte metoder for å lage en bedre grammatikkontroll.

Tekstdata som er tilgjengelig digitalt er stort sett samlet i det nordsamiske korpuset SIKOR (UiT 2018), og bare en liten del er merket opp for grammatikkfeil. Nordsamisk har en relativt ny skriftnormering og det er varierende skriftlig kompetanse blant skribentene. I tillegg har retteverktøy ikke vært tilgjengelig så lenge. Derfor inneholder korpuset mange flere skrive- og grammatikkfeil enn et typisk majoritetsspråkskorpus. Samisk har også en rik morfologi, som betyr at det er mange ordformer og at man trenger enda mer tekst for å dekke alle ordformene.

Dette står i kontrast til store språk der morfologien er relativt enkel, og teksttilfanget er stort og representativt for hele språket. Man fanger lett opp alle ordformer, og man har rik tilgang til språkets syntaks i et slikt teksttilfang. Med et slikt bakgrunnsmateriale man kan lage nevrale nettverk som blir relativt pålitelige fordi ressursene modellen lages på er basert på et allsidig og representativt materiale. For å kompensere for datamangelen har vi derfor laget et nevralt nettverk (maskinlæring) (*NeuSam*) som benytter seg av syntetiske data. Dataene har vi konstruert ved hjelp av regelbasert morfosyntaktisk analyse for å erstatte korrekte former med feilaktige. Etterpå blir dataene filtrert av regelbaserte verktøy - den nordsamiske grammatikkontrollen *GramDivvun*, slik at de syntetiske dataene bare inneholder reelle feil.

2. Problemstilling

Vi tar utgangspunkt i automatisk feilretting i nordsamisk. Den første nordsamiske grammatikkontrollen *GramDivvun* har blitt utviklet siden 2012 og er basert på håndskrevne regler (Wiechetek 2012), og ble offentlig lansert i 2020. Arbeidet til *GramDivvun* er riktignok ikke bare et verktøy for en stor mengde grammatikkfeil på alle områder, dvs. fra ekteordsfeil, til særskrivings- og samsvarsfeil, men også et forskingsresultat for variasjonen i og hyppigheten av nordsamiske grammatikkfeil. Ekteordsfeil er korrekt skrevne ord som er brukt i feil sammenheng. De er vanligvis basert på enten ortografisk eller fonetisk likhet (f.eks. *å* vs. *og*). I denne artikkelen fokuserer vi på retting av samsvarsfeil mellom subjekt og verbal av samme type som i eksempel (1). Samsvarsfeil er en arketypisk grammatikkfeil som er tilstede i mange språk og som krever en analyse av hele setningen. I motsetning til retting av engelske samsvarsfeil i eksempel (1), slik (Ng et al. 2013) tar for seg, er samiske samsvarsfeil langt mer komplekse. Årsaken til dette er at samisk har mange flere verbformer enn engelsk og kombinasjoner av tall (entall, totall, flertall) og person (1.,2.,3.) som må kongruere med verbet. I det samiske eksemplet (2)[1][2] ser man også at det er flere faktorer som må tas hensyn til når subjektet er sammensatt. Subjektet innholder både det personlige pronomenet *mii* i første person flertall og et substantiv i nominativ flertall. Verbet kongruerer med pronomenet og ikke med flertallssubstantivet, det burde derfor være *áigut* isteden for *áigot*. Dette blir synlig på samisk, men ikke på engelsk siden verbformene i *we have* og *they have* er homonyme.

(1) People still *__prefers__ to bear the risk and allow their pets to have maximum freedom.

(2) Mii sámit maid *__áigot__ gullot.
1PL same.3PL også vil.3PL høre.PASS.INF
'Vi samer vil også bli hørt'

Kongruens i nordsamisk gjelder kasus, tall og person, avhengig av kontekst. I nordsamisk er det kongruens mellom subjekt (som er i nominativ) og verb, verb og subjektspredikat, demonstrative pronomener/numeraler og substantiv, og relativpronomen og anafora. (Nickel 1994:s.509ff.)[3]

[1] Alle samiske eksempler er tatt fra SIKOR.

[2] Alle eksemplene følger Leipzig Glossing konvensjonene: http://www.eva.mpg.de/lingua/resources/glossing-rules.php

[3] subjekt og verbal (tall og person - Gal *mun boađán*), verbal og utfyllningspredikativ (Olmmái *lea rikkis*), mellom predikativer (Mus lea *juolgi bávčča*s), objekt og objektpredikativ, relativsetninger (Dat *olmmoš, gii* áigu boahtit.), sammenligning og apposisjoner, (*Máret* lea liikka stuoris go *don* og Oidnet go don *Mihkkala*, min nuoramus *bártni*?

En *kongruensfeil* forutsetter en finitt verbform som ikke samsvarer i tall og person med subjektet som hører til verbalet. Subjektet kan stå enten til venstre eller til høyre for verbalet, og det kan være andre setningsledd mellom subjektet og verbalet. I det følgende eksemplet (3) blir subjektet *makkár váikkuhusat* 'hvilken konsekvenser' og verbalet *ledje* 'var' avbrutt av hovedsetningen *jáhkát don* 'tror du'. I eksempelsetning (4) derimot er det finitte verbet til venstre for subjektet bare en hovedsetning som introduserer en bisetning uten en subjunksjon. Det er *liikojedje* som er verbalet til *mánát* 'barna'. I eksempelsetning (5) er det en relativsetning mellom subjektet *máŋgasat* 'mange' og verbalet *gehččet* 'de ser'.

(3) Makkár váikkuhusat jáhkát don **ledje** dáid lágain sidjiide […]
hvilken konsekvens.NOM.PL tro.2SG 2SG være.PST.3PL disse.GEN lov.LOC.PL de.ILL.PL
'Hvilke konsekvenser tror du disse lovene hadde for dem […]'

(4) *Orui* mánát **liikojedje** oaidnit bihtá.
virke.PST.3SG barn.PL.NOM like.PST.3PL se forestilling.ACC
'Det virket som om barna likte å se forestillingen.'

(5) Sávan máŋgasat, geat eai leat sápmelaččat, **gehččet** dán dokumentára
ønske.1sg mange.PL, som.NOM.PL ikke.3PL være same.NOM.PL, se.PL3 denne.ACC dokumentar.ACC
'Jeg ønsker at mange som ikke er samer, ser denne dokumentaren'

I tillegg til at det kan finnes flere verb som er potensielle verbalkandidater til et subjekt, kan det være ordformer som bare ser ut som finitte verb, men ikke er det. Dette kan skyldes homonymi med finitte verb eller ekteordsfeil. Formen *erret* 'skille' i eksempelsetning (6) er egentlig en ekteordsfeil for adverbet *earret* 'bortsett fra'. Men formen har to verbanalyser, både 1. person flertall og 2. person entall. Det kunne altså tenkes at det er verbalet til *sii* 'de'.

(6) Guossit geat áigot leat sámediggeviesus, ***erret** sii geat áigot leat
gjest.NOM.PL som.NOM.PL vil.3PL være Sametinghus.LOC, skille.1PL;2SG 3PL som vil.3PL være
publikumareálan
publikumsareal.LOC
'Gjestene som vil være i Sametingshuset, bortsett fra de som skal være i publikumsarealet'

Det finnes også systematiske homonymirelasjoner mellom forskjellige former som er presentert i tabell 1. Det er for eksempel noe homonymi mellom perfektum partisipp og første person entall, f.eks. *orron* 'jeg var; har vært'. Alle infinitiver er homonyme med første person presens flertallsverbformer. Infinitiver av ulikestavelses verb og *leat* 'å være' er også homonyme med tredje person flertall. Tredje person presens flertall samsvarer også med andre person preteritum entall ved alle verb bortsett fra *leat* 'være'. Videre samsvarer 1. person presens totall og 3. person preteritum flertall bortsett fra *leat* 'være', ulikestavelsesverb og sammendradde verb. Første person preteritum entall samsvarer med perfektum partisipp-formen ved verb som ender på -ut, f.eks. *gorgŋon* 'jeg har klatret'. I tillegg gjelder denne homonymien for *leat* 'være', ulikestavelses- og sammendradde verb. Noen verb som har endelsen *-ut* har for eksempel passive eller inkoative 3. person entallsformer som er homonyme med aktive 3. person flertallspreteritumsformer, f.eks. *orro* 'hun/han blir boende, de bodde'.

Form	**homonyme former**
INFINITIV	{1. p. flt. / 3. p. flt. presens, 2. p. ent. presens }
PERFEKTUM PARTISIPP	{1. p. ent. preteritum}
1DU PRESENS	{3. p. flt. preteritum }
3. P. FLT. PRETERITUM	{3. p. ent. presens passiv }
BOKTE 'via'	{boktit 'vekke' 3. p. flt. preteritum }
LÁVLU 'sanger'	{lávlut 'synge' 3. p. ent. presens}
…	

Tabell 1: Eksempler på systematiske og idiosynkratiske homonymier

I tillegg til dette finnes det ytterlige idiosynkratiske homonymier, f.eks. *bokte* 'via' som er både en postposisjon og første person totall og tredje person flertall av *boktit* 'vekke'. Andre former er derivasjoner, for eksempel *lávlu* som har en rekke med substantivanalyser ('sanger') og tredje person entall form av *lávlut* 'synge'.

I noen tilfeller er også subjektshomonymi relevant, slik som i setning (7), der tidsskriftet *Dieđut* er homonymt med flertallssubstantivet *dieđut* 'nyheter' og basert på det kunne det tenkes at verbformen må være 3. person flertall.

(7) Dieđut **lea** máŋggadieđalaš čála-ráidu […]
Dieđut.NOM.SG;nyhet.PL være.3SG tverrvitenskapelig skriftserie
'Dieđut er en tverrvitenskapelig skriftserie'

Det er ikke bare homonymi som kan føre til feiltolkinger av setningen. En del syntaktiske fenomen bidrar til utfordringene. En av de største årsakene til unntak er koordinerte subjekt. Mens verbalet *ledje* i eksempelsetning (8) tar hensyn til både første, andre og tredje elementet i koordinasjonen, er det i de fleste tilfellene tillatt med både 3. person entall eller 3. person flertall. Setning (8) koordinerer konkrete personer, i (9) er det derimot mer abstrakte eller uspesifiserte begrep som er koordinert .

(8) Persson, Åberg ja Granberg **ledje** dat golbma buoremusa juohke vuodjimis.
Persson, Åberg og Granberg være.PST.3PL de tre beste hver kjøring.LOC
'Persson, Åberg og Granberg var de tre beste i hver kjøring.'

I eksempelsetning (9) inneholder det koordinerte subjektet *man ollu riggodagat ja ruhta* et flertalls- og et entallssubstantiv. Verbet *manai* er derimot i 3. person entall. Både 3. person entall og 3. person flertall er tillatt.

(9) […] go sii oidne man ollu riggodagat ja ruhta dokko **manai**.
[…] når 3PL se.PST.3PL hvor mye rikdom.NOM.PL og penger.NOM.SG dit gå.PST.3SG
'[…] når de så hvor mye rikdom og penger som gikk dit.'

I setning (10) oppfattes de koordinerte nominalfrasene i subjektet som en logisk enhet, og bare det nærmeste elementet samsvarer med det finitte verbet. Dessuten er samsvar i koordinasjon avhengig av semantisk kategori til substantivene. Ifølge Nickel (1994) «står verbalet i *entall* [hvis subjektsordene er *navn på stoffer*]. […] Hvis subjektsordene er *abstrakte begrep* som nært hører sammen, står verbalet i *entall*.»(s.512)

(10) Sihke jierbmi ja ipmárdus **lea** buorre su iežas adnui.
Både klokhet og forståelse være.3SG bra 3PL.GEN eget bruk.ILL
'Både klokhet og forståelse er bra til sitt bruk.'

Hvis koordinasjonen derimot inneholder et personlig pronomen, er det flertalls- eller totallsformer av samme person som kreves, for eksempel **leimmet** 'vi var' i eksempelsetning (11). Det samme gjelder relativpronomener med et personlig pronomen som antesedent, *midjiide* 'til oss' i eksempelsetning (12), der verbalformen blir 1. person flertall istedenfor 3. person flertall som relativpronomenet.

(11) Oahpaheaddjit **leimmet** fas Isak Johansen, Johan Jernsletten ja mun.
lærer.NOM.PL være.PST.1PL igjen Isak Johansen, Johan Jernsletten og 1SG
'Det var Isak Johansen, Johan Jernsletten og jeg som var lærerne.'

(12) Seamma guoská midjiide geat **bargat** láhččit rámmaeavttuid
samme gjelde.3SG 1PL.ILL som.NOM.3PL jobbe.1PL tilrettelegge.INF rammevilkår.ACC.PL
juohkehačča ovdáneapmái.
enkelte.GEN utvikling.ILL
'Det samme gjelder oss som jobber med å tilrettelegge rammevilkår for den enkeltes utvikling.'

Når verbalet er kopulaverbet *leat* 'være' og det dreier seg om en habitiv eller adverbialkonstruksjon som i (13), så samsvarer det bare med det nærmeste leddet. (Nickel 1994:s.512)[4] I den følgende konstruksjonen (13) er det bare entall som er mulig siden det dreier seg om en konstruksjon med et stedsadverbial i begynnelsen, *dáppe* 'her'.

(13) Mun dieđán dáppe **lea** kultuvra ja árbevierru girkostallat.
1SG vite.1SG her være.3SG kultur.NOM.SG og tradisjon.NOM.SG gå.i.kirken.INF
'Jeg vet at her er det kultur og tradisjon å gå i kirken.'

Visse typer veldig vanlige skrivefeil (ekteordsfeil) kan komplisere søket etter kongruensfeil. I følgende setning (14) er det finitte verbet korrekt. Men i og med at *diehttit* 'å vite' inneholder en skrivefeil (to t-er istedenfor en), blir den mente infinitiven et flertallssubstantiv. Dermed blir det en mulig flertallssubjektskandidat for det finitte verbet, som kunne tolkes som en kongruensfeil - dvs. at det burde være 3. person flertall istedenfor 3. person entall.

(14) Ovddamearkka dihte mo *diehttit **mieđiha** go buohcci vai lea go son
For eksempel hvordan viter.NOM.PL samtykke.3SG QST syk eller være.3SG QST 3SG
duođaid nuppi oaivilis.
egentlig annen mening.LOC
'For eksempel, hvordan skal man vite om den syke samtykker eller om han egentlig har en annen mening.'

En konstruksjon der det kan være vanskelig å finne kongruensfeil, er asymmetriske subjektpredikatskonstruksjoner der subjektet og predikativet ikke har samme tall, som vist i eksempelsetning (15). På språk der subjektet kan være pre- eller postverbalt, slik som i nordsamisk, kan det være vanskelig å identifisere subjektet. (Lorusso et al. 2019) nevner utfordringene i NLP-applikasjoner som for eksempel parsere eller maskinoversetting. Verbalet i italiensk samsvarer med subjektet uavhengig av ordstillinga, på engelsk samsvarer verbalet med den preverbale nominalfrasen som i eksempel (16). (Lorusso et al. 2019)

(15) Davviriikkaid sápmelaččat ***lea** unna minoritehta […]
nordområde.GEN.PL same.NOM.PL være3P.SG liten minoritet.NOM.SG
'Nordens samer er en liten minoritet […]'

(16) a. the pictures are/*is the cause.
b. the cause *are/is the pictures

3. Bakgrunn

3.1. Relatert forskning

Maskinlæringsmetoder som ikke krever lingvistisk ekspertise dominerer per idag moderne språkteknologi (f.eks. (Chollampatt and Ng 2018, Boyd 2018)). Fokuset i maskinlæring har vært på maskinoversetting og andre typer verktøy. Maskinlærte stavekontroller skiller vanligvis ikke på vanlige skrivefeil og grammatiske feil. I det siste har store datamengder ført til at resultatene har bedret seg noe og medført at man har kunnet laget mer avanserte grammatiske verktøy som blir brukt av et bredt publikum.

Det er få eksempler på grammatikkontroller som er basert på nevrale nettverk som er i daglig bruk og er veldokumentert. Noen av de mest populære systemene i bruk er fortsatt regelbasert, slik som *LanguageTool*[5] (basert på åpen kildekode). *Grammarly*[6], som er lukket programvare, bruker maskinlæringsmetoder til en

[4] «Hvis predikativet består av flere sidestilte ord i nominativ, så er det vanligvis samsvar i tall mellom verbalet og det ordet i predikativet som står nærmest. Dette gjelder setninger med habitiv eller adverbial i nominatdelen» (p.512)

[5] https://languagetool.org

[6] https://grammarly.com

viss grad [7].

På begynnelsen av 90-tallet introduserte Fred Karlsson konseptet føringsgrammatikk (Constraint Grammar). Denne teknologien har produsert gode tekstprosesseringsverktøy, bl.a. grammatikkontroller, som har blitt godt mottatt og brukt i mange språksamfunn (Arppe 2000, Birn 2000, Hagen and Lane 2001). I *GiellaLT*-infrastrukturen blir det utviklet føringsgrammatikker der lingvisten har kontroll over hvordan grammatikkontrollene fungerer og hvilke problem de skal løse. Det er ikke bare tekniske årsaker for metodevalget. Kunnskapsøkning om grammatikken til det språket som jobbes med, kvalitetssikring og kontrollerbarhet (grammatikkontrollen gjør det den skal gjøre også ifølge menneskelige standard) ligger bak preferansen om å jobbe regelbasert.

3.2. Våre ressurser

I dette eksperimentet bruker vi *GiellaLT*-infrastrukturen[8] for å lage digitale grammatikker og leksikon og for å lage verktøy som bruker disse grammatikkene og leksikonene (Moshagen et al. 2014). Infrastrukturen er bygd opp slik at verktøyene (tastatur, stavekontroller, etc.) er laget på samme måte for alle språkene, og skiller på denne måten mellom språkspesifikke data og språkuavhengige metoder. *GiellaLT* har for tiden repositorier for 136 forskjellige språk – for det meste (sirkumpolære) minoritetsspråk eller andre mindre språk. Denne artikkelen bygger på den nordsamiske delen av infrastrukturen[9] og er et eksperiment for å eventuelt introdusere nye nevrale metoder til det språkuavhengige byggesystemet.

For å evaluere og trene den nevrale modellen bruker vi SIKOR. SIKOR inneholder ca. 39M ord og består av to korpora: *GT-Bound*[10] (tekster som er dekket av opphavsrett og som er tilgjengelig på forespørsel) og *GT-Free*[11] (tekster som er offentlig tilgjengelig). For å evaluere resultater for både den regelbaserte og den nevrale modellen, bruker vi et gullkorpus på ca 406 000 ord som er en del av *GT-Free* og *GT-Bound* og som er oppmerket med mange forskjellige feiltyper.

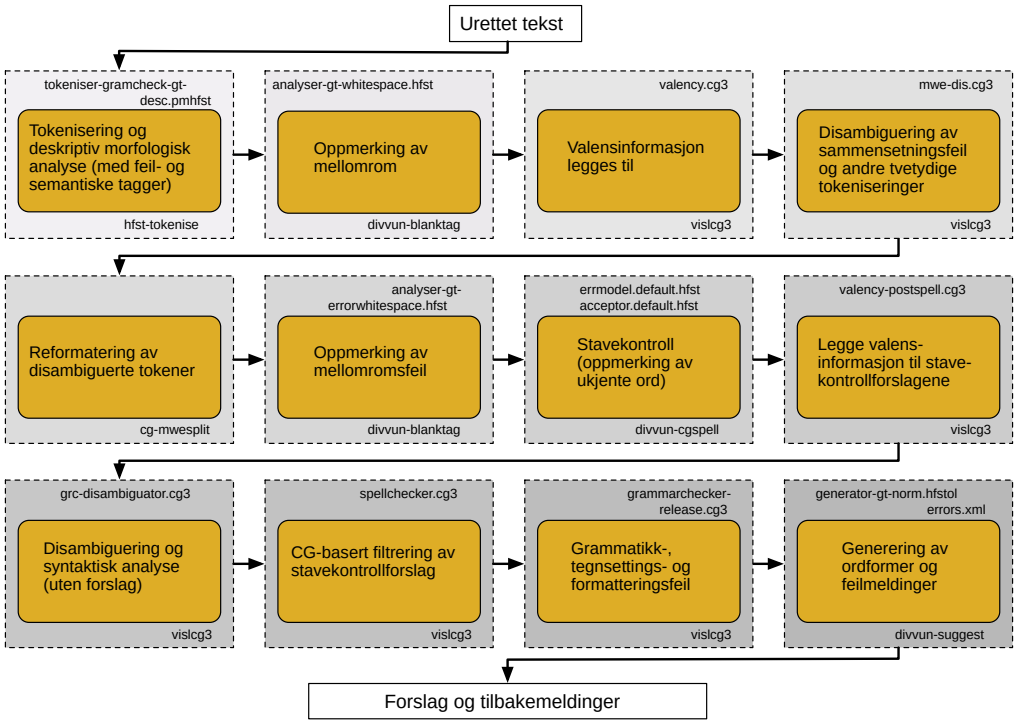
4. Metodevalg

4.1. Regelbasert metode (GramDivvun)

Kongruensfeilretting ved hjelp av håndskrevne regler er basert på endelige tilstandsautomater (FST) (Beesley and Karttunen 2003, Pirinen and Lindén 2014) og føringsgrammatikker (Constraint Grammar) (Karlsson 1990). Den nordsamiske regelbaserte grammatikkontrollen *GramDivvun* retter både skrive- og mange grammatikkfeil i tillegg til tegnsettings- og formateringsfeil. *GramDivvun* er bl.a. tilgjengelig som en plugin for Microsoft Office og Google Docs[12] og er åpen kildekode.[13] Den inkluderer bl.a. en nyere versjon av stavekontrollen fra 2007[14], cf. also (Gaup et al. 2006), og seks føringsgrammatikkmoduler, se figur 1.

Kongruensfeilretting foregår i 'grammarchecker-release.cg3'-modulen. 45 regler legger til en samsvarsfeiltag til verbformen som skal rettes. Hver kombinasjon av person og tall har et eget regelsett som vanligvis består av forskjellige regler for pre- og postverbal subjektsposisjon. I tillegg er det spesifikke regler for passivkonstruksjoner, negasjonskontekster, relativsetninger, kopula, adposisjoner og koordinerte subjekter. Regelsettet for pronominale førstepersonsflertallskontekster er litt mer komplekst siden formen *mii* er

[7] https://www.grammarly.com/blog/engineering/grammarly-nlp-building-future-communication/
[8] https://giellalt.github.io
[9] https://github.com/giellalt/lang-sme
[10] https://gtsvn.uit.no/boundcorpus/orig/sme/
[11] https://gtsvn.uit.no/freecorpus/orig/sme/
[12] https://divvun.no/no/korrektur/gramcheck.html
[13] den presise versjonen som er brukt i eksperimentet finnes her for reproduksjon: https://github.com/giellalt/lang-sme/releases/tag/experiment-2022-03-30 se også https://github.com/giellalt/giella-core/releases/tag/experiment-2022-03-30 og https://github.com/giellalt/giella-shared/releases/tag/experiment-2022-03-30
[14] http://divvun.no/korrektur/korrektur.html



Figur 1: Modulær struktur av *GramDivvun*

homonymt og kan være både 1. person flertall ('vi') og et spørrepronomen i 3. person entall ('hva').

Reglene som legger til feiltaggene til en feilaktig verbform har følgende format (forenklet) og følger 'Constraint Grammar'-formalismen. Regelen nedenfor (som er en av 48) går ut i fra en 3. person entalls-høyrekontekst.

```
ADD (&kongruensfeiltag)
TARGET finitte verbformer bortsett fra konnegativ/negasjonsverb
IF i høyre kontekst det er et personlig pronomen i 3. person entall
   som ikke inneholder en feil
   det ikke finnes et annet verb i 3. person entall til høyre for det og
   verbet har ingen 3. person entalls-/perf.part.-/konnegativ-/adverbslesing
   verbet har ingen 3. person flertallslesing med et koordinert subjekt til høyre
   [...];
```

4.2. Nevral metode (NeuSam)

4.2.1. Datagenerering (syntetiske feil)

Nevrale nettverk krever en stor mengde av parallelt korpus mellom korrekte og feilaktige setninger. Siden det kan ta flere år å bygge et slikt korpus, er det vanlig å generere et feilkorpus. Ulempen med et generert feilkorpus er at det innebærer en risiko for at feilfordelingen ikke er representativ eller at feilene kanskje ikke er feil. Dataene vi bruker i dette eksperimentet kommer fra SIKOR, og blir viderebehandlet med skript som genererer grammatikkfeil. Vi analyserer korpuset med *GramDivvun* og fjerner setninger med feil, for å

deretter introdusere feil ved å forandre på ordformene i dette materialet. Utfordringene med strategien har vært:

- For å ikke generere den samme formen som den feilaktige, har vi filtrert bort de introduserte formene som er homonyme (*leat* 'vi er', *leat* 'du er').
- Siden datamengden øker eksponensielt om vi erstatter en form med mange andre, spesielt når det er flere verb i setningen, har vi valgt å bare introdusere en feil av gangen i setningen, istedenfor å kombinere alle variantene.

Den korrekte setningen (17) som inneholder et 3. person entallssubjekt og en 3. person entallsverbform kan brukes for å generere opptil 8 setninger med en syntetisk feil (eksempel (18)). Dette gjøres ved å erstatte den korrekte verbformen med forskjellige feilaktige former som er forskjellig i person og tall (som ikke er homonyme med den rette formen).

(17) Son **doarjjui** áinnas unnit *giliid.
3SG støtte.PST.3SG selvfølgelig mindre språk.ACC.PL
'Hun støttet selvfølgelig mindre språk.'

(18) a. Son **dorjot** áinnas unnit giliid.
b. Son **doarjjuiga** áinnas unnit giliid.

Vi brukte et skript[15] som leser gjennom hver setning i korpuset, og for hver analyse erstatter skriptet verbformen som kan ha kongruens med et subjekt med andre verbformer som ikke har kongruens med subjektet. En oversikt av erstatninger som ble gjort vises i tabell 2. I den første gruppen valgte vi bare et verb og erstattet det med andre former (f.eks tar vi et verb i første person entall og erstatter det med 2. person entall og 3. person entall, og alle totalls- og flertallsformene). I den andre gruppen genererte vi frekvente grammatikkfeil, som tilsvarer feil basert på vår erfaring med korpussøk. Ordene i den andre gruppen har også en begrensing av fonologisk form, f.eks. IND PRS PL3[16] til IMPRT PL2-feil er en feil som oppstår i likestavelsesverb. Etterpå filtrerte vi de genererte setningene med *GramDivvun* igjen, slik at vi bare satt igjen med setninger *GramDivvun* anså for å være feil. Resultatet er at flesteparten av de syntetiske feilene som vi introduserte, hhv. 94.5% og 86.4%, ikke ble merket som feil av *GramDivvun*, antakeligvis fordi de er korrekte med formen som ble erstattet. Dette er ikke uvanlig med tanke på at setninger uten subjekt kan ha korrekte verbformer i alle slags person-tall kombinasjoner. Vi valgte å bruke *GramDivvun* for å filtrere setningene etter at vi ved en manuell gjennomgang oppdaget at feilkorpuset som ble generert for å trene *NeuSam* inneholdt mange setninger som var korrekte. Siden *GramDivvun* tidligere viste seg å ha god presisjon valgte vi å redusere feilkilden ved å bare trene *NeuSam* med setninger *GramDivvun* anser som feil.

4.2.2. Trening og testing

Vi har brukt OpenNMT-py (Klein et al. 2017) for eksperimenteringen med nevrale nettverk. Vi fulgte metoden som er beskrevet i OpenNMT-py sin 'tutorial'[17] med standardparametrene.

90 % av dataene vi samlet i stegene ovenfor ble brukt for å trene modellene. Vi reformaterte dataene våre slik at de ble tolket som en bokstavbasert modell. Dette gjorde vi for å unngå OpenNMTs automatiske tokenisering. Disse parametrene vises også i tabellen 3.[18]. Trening av modellen ble gjort på en GPU-supercomputer fra «*UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway*». Det tok i gjennomsnitt fem timer å generere hver treningsmodell.

[15] https://gtsvn.uit.no/hybrid_gramcheck

[16] vi bruker *GiellaLT* sine analysetaggar som er dokumentert her: https://giellalt.github.io/lang-sme/docu-mini-smi-grammartags.html

[17] https://opennmt.net/OpenNMT-py/quickstart.html

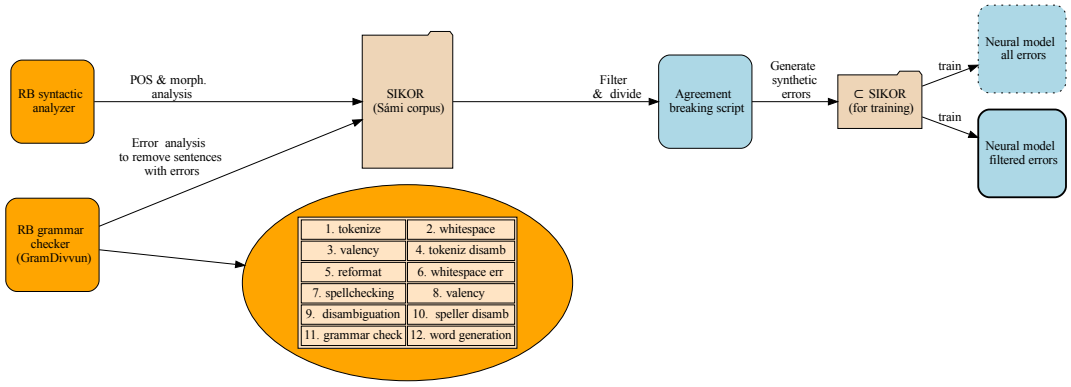
[18] Vi inkluderer hele konfigurasjonen av opennmt-py og skript til trening i https://gtsvn.uit.no/hybrid_gramcheck ved publisering

Analyse →	**Syntetisk**
(V) SG1	{Sg2, Sg3, Du1, Du2, Du3, Pl1, Pl2, Pl3}
(V) SG2	{Sg1, Sg3, Du1, Du2, Du3, Pl2}
(V) SG3	{Sg1, Sg2, Du1, Du2, Du3, Pl1, Pl2, Pl3}
(V) DU1	{Sg1, Sg2, Sg3, Du2, Du3, Pl1, Pl2}
(V) DU2	{Sg1, Sg2, Sg3, Du1, Du3, Pl1, Pl2, Pl3}
(V) DU3	{Sg1, Sg2, Du1, Du2, Pl1, Pl2, Pl3}
(V) PL1	{Sg1, Sg3, Du1, Du2, Du3, Pl2}
(V) PL2	{Sg1, Sg2, Sg3, Du1, Du2, Du3, Pl1, Pl3}
(V) PL3	{Sg1, Du2, Du3, Pl1, Pl2}
(V) IND PRS PL3	IMPRT PL2
(DER/PASS V)	IMPRT DU2
Ind Prs Sg3	
(V) IND PRS SG3	IND PRT PL3

Tabell 2: Erstatninger for å generere grammatikkfeil; kontekst er i parentes.

Parameter	**Verdi**
`train steg`	`100,000`
`valid steg`	`10,000`
`vocab størrelse`	`50,000`
`seed`	`3,435`
`encoder type`	`brnn`

Tabell 3: Parametre gitt til OpenNMT



Figur 2: Et diagram av *NeuSam* og treningsprosessen

Vi har generert to forskjellige nevrale modeller med forskjellige datasett: en med et større datasett der vi bruker alle syntetisk genererte setninger som omtalt i seksjonen 4.2.1. I den andre lager vi en modell basert på setninger som etter syntetisk feilgenerering blir filtrert gjennom *GramDivvun*. Input til testene av de nevrale modellene er den tiendelen av vårt genererte korpus som ikke har blitt brukt i treningen av modellene, og testen vi gjør er å sjekke hvor stor del av dette testsettet som blir merket som feil. Formelen for nøyaktighet er ganske enkel: nøyaktighet $= \frac{\text{korrekte}}{\text{alle}}$ der *korrekte* er antall setninger som modellen anser for å inneholde feil, *alle* er antall setninger i testsettet.

I tabell 4 ser vi at modellen basert på filtrerte setninger er mer nøyaktig. Den større modellen har 9 % dårligere resultat enn den mindre modellen. Det betyr at modellen basert på ufilterte setninger egentlig har lært å fikse feil deler av eller ikke fikser alle feil i nesten 1 av 10 setninger med syntetiske feil.

Modell	Nøyaktighet
Stor	25 %
Filtrert	37 %

Tabell 4: Nøyaktighet av nevrale modeller

Modell	Presisjon	Dekning	F-Score
GramDivvun	78.50 %	43.75 %	56.19
NeuSam	27.01 %	8.21 %	12.61

Tabell 5: Evaluering av den regelbaserte og maskinlæringsmodellen

5. Resultater

Vi har evaluert *NeuSam* og *GramDivvun* på det oppmerkede korpuset på 406 000 ord som er en del av SIKOR. Korpuset består av mange administrative og nyhetstekster, litt skjønnlitteratur og en del L2-tekster som ble samlet inn for spesielle formål. Oppmerkingen fulgte opprinnelig noen retningslinjer for skrivefeil og fonologiske prosesser, og den har skjedd over et lengre tidsrom, ca. 15 år. Etterhvert ble oppmerkingen utvidet og tilpasset grammatikkontroll og måten evalueringsskriptet er istand til å kjenne igjen disse feilene på. Vi følger prinsippet om at bare det som blir rettet blir merket opp og ikke konteksten for å se feilen. Vi oppdaget en del inkonsekvent oppmerking som vi rettet under dette arbeidet. Dette skyldes også at grammatikkontrollprogrammet kom mange år etter at korpusoppmerkingen startet. Grammatisk feilkategorisering var ikke helt utarbeidet på det tidspunktet og man kunne ikke sjekke mot et dataprogram som krever konsekvent oppmerking.

Vi ønsket å sammenligne presisjon og dekning og sjekke om *NeuSam* retter feil som *GramDivvun* ikke oppdager. Tabell 5 viser at *GramDivvun* er betydelig bedre enn *NeuSam* på å finne kongruensfeil. Man ser også at mange av korreksjonene til *NeuSam* ikke har noen lingvistisk forklaring, mens mange av korreksjonene til *GramDivvun* kan være nyttig for brukeren i og med at de viser til en annen feil i setningen.

I eksempelsetning (19) blir verbformen *livčče* 'de skulle' rettet til entall *livčči* 'hun/han skulle' fordi subjektet *Maáŋga mearraolbmáidgirku* inneholder et tallord med en skrivefeil. Dermed oppfattes bare entallssubstantivet *mearraolbmáidgirku* som subjekt, og flertallsbetydninga blir tapt. Dette regnes som en falsk positiv i evalueringen, men grammatikkontrollen har 'tenkt' rett ut i fra den informasjonen som er tilgjengelig (altså før skrivefeilen blir rettet).

(19) Maáŋga mearraolbmáidgirku **livčče** vuollebáhcagiin šaddan rahčat […]
mange sjømannskirke.NOM;GEN.SG ville.POT.3PL underskudd.COM.PL bli.PASTP kjempe.INF
'Mange sjømannskirker ville kjempet med underskudd […]'

Falske negativer er det flest av i koordinasjon med to eller flere substantiv, der det finitte verbet skal være i entall isteden for flertall. Et eksempel er (20), der tredje person flertallsformen *leat* skal rettes til tredje person entallsformen *lea*. Kongruens i koordinasjon er avhengig av flere faktorer, blant annet semantisk tilhørighet, syntaks (kopulakonstruksjoner og adverbialkonstruksjoner behandles forskjellig fra andre) og pragmatikk (er den introduserte entiteten kjent?). Grunnen til at feilen ikke blir oppdaget er at vi ennå ikke laget en regel som retter fra flertalls- til entallsverb i koordinasjon.

(20) Álggahanvahkku prográmmas **leat** almmolaš rahpanbeaivi,
startuken.GEN program.LOC være.3SG offisiell åpningsdag,
diehtojuohkin Sámi allaskuvlla birra, fáddarortnet ođđa studeanttaide.
informasjon Samisk høgskole om, fadderordningen nye student.ILL.PL.
'I startukens program inngår offisiell åpningsdag, informasjon om Samisk høgskole, fadderordningen for nye studenter.'

At *NeuSam* ikke finner flesteparten av de oppmerkede feilene, skyldes sannsynligvis at treningsmaterialet ikke er representativt nok. Et annet problem er at når rettingen går galt, blir rettelsen helt uforståelig. Et eksempel er at samme ordrekkefølge blir repetert uendelig mange ganger *johtá guovllus sahtá guovllus sahtá guovllus sahtá guovllus sahtá guovll...*. Dette lar seg fikse ved å endre på lengderestriksjoner for setninger,

men følgen er at man ikke kan rette lengre setninger.

Den større modellen gir følgende feilaktige resultat for eksempel (21-a): Istedenfor å bare rette verbformen *logat* 'du leser' til *lohká* 'hun/han leser' blir setningen rettet til (21-b), dvs. *NeuSam* tar bort hele setningen *logan dál oppalaččat* uten at dette skulle være lingvistisk fundert.

(21) a. In dovdda dán áššši, *logan dál oppalaččat,* **logat** Sámedikki presideanta Egil Olli.
b. In dovdda dán áššši, **lohká** Sámedikki presideanta Egil Olli.
c. In dovdda dán áššši, logan dál oppalaččat, **lohká** Sámedikki presideanta Egil Olli.

NeuSam produserer også noen falske positiver, f.eks. i (22) blir *šaddet* rettet til *šaddá* (3Pl>3Sg), men det burde ikke rettes siden *stuorát doalut* er et flertallssubjekt.

(22) Duogážin manne heastasearvi lea fárus doaluin, lea danin vai
bakgrunn.ess hvorfor hesteforening være.3sg med arrangement.loc.pl, være.3sg derfor at
šaddet stuorát doalut […]
bli.3pl stor.comp arrangement.nom.pl
'Bakgrunnen for at hesteforeningen er med i arrangementet, er at det blir et større arrangement'

6. Konklusjon

I denne artikkelen laget vi to maskinlæringsmodeller for å rette kongruensfeil mellom subjekt og verbal i nordsamisk. Parallelt med dette utviklet vi et regelsett for slike feil i *GramDivvun*, den eksisterende regelbaserte grammatikkontrollen. Vi ville sammenligne resultatene for maskinlæring og regelbasert metode, både for å få mer klarhet i hvilken metode som bør foretrekkes for dette formålet og for å se om systemene har styrker på forskjellige områder og kan kombineres til en hybrid grammatikkontroll. Vi ville også forsøke å avdekke myten om at maskinlæring blir billigere enn regelbaserte metoder, og det mener vi at vi har gjort ved å tydeliggjøre at det å generere treningsdata må regnes inn i de faktiske kostnadene til metoden. For å lage et feiloppmerket treningskorpus for *NeuSam* brukte vi den regelbaserte modellen *GramDivvun* for å rydde korpuset for støy. Dette var nødvendig for å etterpå kunne introdusere syntetiske feil. Uten denne filteringen blir nøyaktigheten til *NeuSam* 12 prosentpoeng verre. Det at den regelbaserte modellen blir brukt for å automatisk generere data viser at korpuset ikke blir gratis.

Vår hypotese – at regelbaserte metoder kan kompensere for mangel av data, også for maskinlæringsmodeller – har vist seg å ikke holde stikk når det gjelder retting av globale grammatikkfeil. Evalueringen på et ekte korpus (dvs. med ekte feil i en naturlig distribusjon) i tabell 4 viser at for den regelbaserte modellen er presisjonen nesten tre ganger bedre og dekningen fem ganger bedre enn for den maskinlæringsbaserte modellen. *GramDivvun* presterer så bra (79% presisjon) at vi har en modell som er til nytte for språkbrukere i og med at mengden på de falske alarmene er relativt lavt. *NeuSam* derimot gjør det såpass dårlig på et ekte korpus, med en presisjon på bare 27% (på testsettet var resultatene tre ganger bedre), at det ikke kan brukes for å lage en hybrid grammatikkontroll for kongruensfeil. Det taler for at det syntetiske feilkorpuset kanskje ikke er representativt nok til å være et realistisk feilkorpus. I tillegg er det å introdusere ekte kongruensfeil en oppgave som krever mer enn enkle erstatninger og en enkel kontekstanalyse. Mange kontekster tillater flere former uten at disse er feil. Det å introdusere kongruensfeil kan anses som en oppgave som er minst like vanskelig som, om ikke vanskeligere enn, selve feilfinningen. Dvs. at vi trenger et verktøy som er like bra som den regelbaserte grammatikkontrollen for å lage et korpus for en maskinlæringsbasert grammatikkontroll. Mens maskinlæringsmetoder fungerer for mer lokale feil som for eksempel sammensettingsfeil, er det for krevende å lage feilkorpus for mer avanserte feil. Dette gir et bra utgangspunkt for framtidig forskning, men med de nåværende ressursene synes ikke maskinlæring å være den mest lovende metoden for å lage grammatikkontroller. Den regelbaserte metoden er fortsatt den som gir best resultat på dette området.

Godord

Modelleringen av de nevrale nettverkene har blitt utført på maskinene til UNINETT Sigma2.

Referanser

Arppe, Antti. 2000. Developing a grammar checker for Swedish. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NoDaLiDa 1999)*, edited by Torbjørn Nordgård, pp. 13–27. Department of Linguistics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

Beesley, Kenneth R and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI publications.

Birn, Juhani. 2000. Detecting grammar errors with Lingsoft's Swedish grammar checker. In *Proceedings of the 12th Nordic Conference of Computational Linguistics (NoDaLiDa 1999)*, edited by Torbjørn Nordgård, pp. 28–40. Department of Linguistics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

Boyd, Adriane. 2018. Using wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pp. 79–84. https://doi.org//10.18653/v1/W18-6111.

Chollampatt, Shamil and Hwee Tou Ng. 2018. A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

Dahlmeier, Daniel, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31. Association for Computational Linguistics, Atlanta, Georgia.

Gaup, Børre, Sjur Moshagen, Thomas Omma, Maaren Palismaa, Tomi Pieski, and Trond Trosterud. 2006. From Xerox to Aspell: A first prototype of a North Sámi speller based on TWOL technology. In *Finite-State Methods and Natural Language Processing*, edited by Anssi Yli-Jyrä, Lauri Karttunen, and Juhani Karhumäki, pp. 306–307. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11780885_37.

Hagen, Kristin and Pia Lane. 2001. "det er fort gjort og skrive feil." en presentasjon av en automatisk grammatikkontroll for bokmål pp. 93–102.

Karlsson, Fred. 1990. Constraint grammar as a framework for parsing unrestricted text. In *Proceedings of the 13th International Conference of Computational Linguistics*, edited by H. Karlgren, vol. 3, pp. 168–173. Helsinki. https://doi.org/10.3115/991146.99117.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72. Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/P17-4012.

Lorusso, Paolo, Matteo Greco, Cristiano Chesi, and Andrea Moro. 2019. Asymmetries in extraction from nominal copular sentences: a challenging case study for nlp tools. In *Proceedings of the Sixth Italian Conference on Computational Linguistics Bari (CliC-it 2019)*.

Miłkowski, Marcin. 2007. Automated building of error corpora of polish. *Corpus Linguistics, Computer Tools, and Applications – State of the Art. PALC* pp. 631–639.

Moshagen, Sjur, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC*, pp. 71–77.

Ng, Hwee Tou, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–12. Association for Computational Linguistics, Sofia, Bulgaria.

Nickel, Klaus Peter. 1994. *Samisk grammatikk*. Davvi Girji, Kárášjohka, second edn.

Pirinen, Tommi A. and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent*

Text Processing - Volume 8404, CICLing 2014, pp. 519–532. Springer-Verlag, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-54903-8_43.

Simons, Gary F. and Charles D. Fennig (eds.) . 2018. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, twenty-first edn.

UiT. 2018. SIKOR uit norges arktiske universitets og det norske sametingets samiske tekstsamling, versjon 06.11.2018. http://gtweb.uit.no/korp. Accessed: 2018-11-06.

Wiechetek, Linda. 2012. Constraint Grammar based correction of grammatical errors for North Sámi. In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8/AFLAT 2012)*, edited by G. De Pauw, G-M de Schryver, M.L. Forcada, K. Sarasola, F.M. Tyers, and P.W. Wagacha, pp. 35–40. European Language Resources Association (ELRA), Istanbul, Turkey.

Wiechetek, Linda, Flammie Pirinen, Mika Hämäläinen, and Chiara Argese. 2021. Rules ruling neural networks - neural vs. rule-based grammar checking for a low resource language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pp. 1526–1535. INCOMA Ltd., Held Online. https://doi.org/https://doi.org/10.26615/978-954-452-072-4_171.

Čalbmi čalmmis ja suoldnečalmmit suoidnečalmmis

Sámegielaid singulatiivvat

Jussi Ylikoski
Oulu universitehta ja Sámi allaskuvla

Abstract

North Saami *čalbmi* ‘eye’ (< Proto-Uralic *ćilmä) has cognates in all Uralic languages, and everywhere they refer to the visual organs of humans and animals. However, scholars have barely paid attention to the grammatical functions of čalbmi in compound-like formations such as suoldnečalbmi “dew eye”, suoidnečalbmi “grass eye”, varračalbmi “blood eye”, jiekŋačalbmi “ice eye”, vuoktačalbmi “hair eye” and muorječalbmi “berry eye”. This article examines such expressions as so-called singulatives – grammatical means for individuating a single referent from a group or mass (i.e., ‘a single drop of dew’, ‘a single blade of grass’, ‘a single drop of blood’, ‘a single crystal of ice’, ‘a single human hair’ and ‘a single berry’). The article mainly discusses morphological, syntactic and semantic features of singulatives in North Saami and other present-day Saami languages, but comparable singulatives in Khanty, Mansi and Samoyed languages as well as in Hungarian suggest that singulative expressions such as *weri-ćilmä ‘a single drop of blood’, *jäŋi-ćilmä ‘a single crystal of ice; hailstone’ and *mẹrja-ćilmä ‘a single berry’ can, in principle, be reconstructed all the way back to Proto-Uralic.

Keywords: Saami languages, singulatives, number

1. Álggahus

Sámegielaid *čalbmi* – mainna dán čállosis čujuhan maiddái omd. sániide *tjelmie*, *tjalmme*, *čalme*, *čâ'lmm* ja *чалльм* – lea okta dain boarrásamos sániin, maid sámegielat leat árben gitta urálalaš vuođđogielas, mas sánis **ćilmä* leamaš seamma universála mearkkašupmi go dálá sámegielain ja dadjat juo visot urálalaš gielain, main dát sátni lea seailluhan iežas vuođđomearkkašumi, vaikko hápmi dieđus lea rievdan. Dán čállosis in dattetge olusge guorahala dáid sániid primára mearkkašumi muhto baicce dan, makkár rolla *čalmmis* lea sámegielaid giellaoahpas. Čuoččuhan, ahte *čalbmi* lea eanet go sátni: dan sáhttit atnit maiddái muhtunlágan giellaoahpalaš elemeantan, dihto láhkai measta sojahangeažusin (*-čalbmi*), jus mearridit guorahallat áššis dan giellaoahpalaš geahččanguovllus mii lea dán čállosa vuođđun.

Dán čállosis guorahalan *-čalmmi* dihtolágan numerus- dahjege lohkodovddaldahkan, man sáhttá gohčodit *singulatiivan*. Sáhkan ii leat *singulára* dahjege ovttaidlohku, muhto baicce *singulatiiva*, man mearkkašumi čilgen dárkileappot kapihttalis 2. Kapihttalis 3 ges ovdanbuvttán konkrehta ovdamearkkaid sámegiela singulatiivvain ja guorahalan daid earenoamážit semantihka geahččanguovllus. Loahpas, kapihttalis 4, digaštalan *čalbmi*-singulatiivvaid mearkkašumi sámegielaide – daid giellaoahpaide – ja áinnas sámegielaid dutkiide ja geavaheddjiide. Oanehaččat daddjon guorahalan sátnehámiid dego *suoldnečalmmit* ja *suoidnečalbmi*, maid refereanttat oidnojit govas 1, mii ii mange láhkai muital maidege dakkár čalmmiid birra maiguin mii dán gova oaidnit.

 Nordlyd 46.1: 299–307, *Morfologi, målstrev og maskinar*: *Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, maid Lene Antonsen, Sjur Nørstebø Moshagen ja Øystein A. Vangsnes leat doaimmahan. UiT Norgga árktalaš universitehta lea almmuheaddji. http://septentrio.uit.no/index.php/nordlyd
https://doi.org/10.7557/12.6304

Govva 1: Suoldnečalmmit suoidnečalmmis. Gáldu: Wikimedia (https://commons.wikimedia.org/wiki/File:Fullsizeoutput_3d5Morgen_dew.jpg).

Čuoččuhan dán čállosis, ahte áidna oktasaš semantihkalaš sárggus sátnehámiin *suoldnečalbmi*, *suoidnečalbmi* ja omd. *sáttočalbmi* lea mearkkašupmi ’okta’ dahje ’muhtun’.

2. Dutkama duogáš ja ulbmilat

Dán čállosa fáddán lea sámegielaid morfema *-čalbmi* singulatiivva dovddaldahkan. Tearbma ja doaba *singulatiiva* boahtá dálá álbmogiidgaskasaš gielladiehtaga bokte kelttalaš gielaid giellaoahpain, main leat hállan ng. singulatiivvaid birra juo Zeuss (1853: 299–301) rájes. Dainna oaivvildit giellaoahpalaš morfemaid, maid mearkkašupmi lea seamma go singuláras, ovttaidlogus, muhto dat lea namalassii morfema ja iige nullamorfema. Kymrigielas dát mearkkaša omd. dan, ahte sánit *adar* ’lottit’, *plant* ’mánát’ ja *coed* ’vuovdi; muorat’ čujuhit máŋgga referentii, muhto čujuhan dihte namalassii ovtta loddái, ovtta mánnái ja ovtta murrii fertet lasihit daidda singulatiivagehčosa *-yn* dahje *-en*: *ader**yn*** mearkkaša ’(okta) loddi’, *plent**yn*** ’(okta) mánná’ ja *coed**en*** ges ’(okta) muorra’.

Eurohpás singulatiivvat leat adnon lagamustá kelttalaš fenomenan, maidda gávdnojit gal vástagat Eurohpá máttimus gielaide gullevaš maltagielas (omd. *zarbun* ’gápmagat’, *ħut* ’guolit’ ja *lariṅġ* ’appelsiinnat’, muhto singulatiivvat *zarbuna* ’gáma’, *ħuta* ’guolli’ ja *lariṅġa* ’appelsiidna’) ja de ain guhkkelabbos dan lagas fuolkegielas arábagielas ja duoppil dáppil miehtá máilmmi. *Singulatiiva*-doaba lea dattetge oalle ođas dábálaš gielladiehtagis, ja giellatypologalaš dutkan dán fáttás lea easka álgán (omd. Haspelmath ja Karjus 2017; Grimm 2018; Dali ja Mathieu 2021). Danin ii leat imaš, ahte singulatiivvaid dutkit eai leat olusge gidden fuomášumi urálalaš gielaid dego sámegiela potentiála singulatiivvaide, dasgo olles doaba lea leamaš amas min giellaoahppaárbevirrui ja dan suopmelaš ja skandinávalaš duogážii.

Dál go buvttán ođđa tearpma sámegiela dutkamii, de lea dehálaš deattuhit ahte *singulatiiva* lea vuosttažettiin morfologalaš tearbma. Dat čujuha dasa, ahte dihto substantiivvas lea dihto earenoamáš hápmi, vaikko dan numerála mearkkašupmi ii spiehkkas dábálaš, nu gohčoduvvon rehkenasttehahtti substantiivvaid (eng. *countable noun*) ovttaidlogu mearkkašumis. Singulatiivva funkšuvdnan lea dattetge earuhit ovttaskas elemeantta, bihtáža dahje partihkkalačča mii muđui gullá stuorát homogena jovkui dahje ávdnasii, mas eai leat čielga ráját dahje eai ráját ollege. Vaikko ovddit ovdamearkkat dego sánit mearkka-šumiiguin ’loddi’, ’guolli’ ja ’appelsiidna’ eai goit sámegiela perspektiivvas buvtte millii prototiippalaš ávnnassániid, de ovdamearkkat leatge gielain main singulatiiva lea mearkkašahtti giellaoahpalaš kategoriija. Máilmmi gielain dábálut orru dattetge leamen dat, ahte substantiivvat main leat earenoamáš singulatiivahámit čujuhit golgosiidda (omd. *čáhci*, *varra*) dahje eará dábálaččat juogekeahtes ávdnasiidda (*jiekŋa*, *muohta*, *sáddu*, *suoldni*) dahje uhcit eanet kollektiiva joavkkuide (*guolga*, *hilla*, *luomi*, *suoidni*).

Aiddo dánlágan substantiivvaid oktavuođas geavahit sámegielain elemeantta *-čalbmi* – muhtumin maiddái dalle go ovttageardán ovttaidlogu hápmi orošii prinsihpas leamen doarvái, muhto ii dattetge leat. Omd. sámegiela sátni *vuokta* adnojuvvo dábálaččat plurale tantum -sátnin mii gullo ja oidno ovttaidlogu hápmin measta dušše beare goallossániin dego *vuoktačuohppi* ja *vuoktabáddi*, muhto muđui hállat *vuovttaid* birra. Muhto dalle go hállat ovttaskas *vuoktačalmmi* birra, geavahit dábálaččat man nu sivas elemeantta

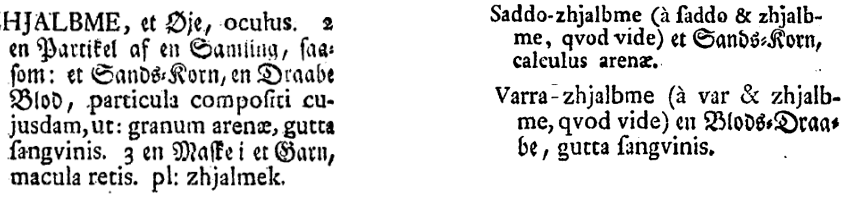
-čalbmi, vaikko dán elemeantta mearkkašumis illá lea gaskavuohta oaidninorgánaide. Dan dihte evttohan dán čállosis, ahte maiddái sámegiela *-čalbmi* sáhtášii dulkojuvvot seammasullasaš singulatiivva dovddaldahkan go kymrigiela *-yn* ja maltagiela *-a* čuovvovaš sojahanpárain. Namahus *kollektiiva* čujuha dán oktavuođas substantiivvaid vuođđohámiide, main ii leat ovttaidlogu mearkkašupmi muhto baicce juoga man sáhttá karakteriseret kollektiivan:

	Kollektiiva		Singulára	
Kymrigiella	*blew*	'vuovttat'	*blew-yn*	'vuoktačalbmi'
	gwellt	'suoidni'	*gwellt-yn*	'suoidnečalbmi'
	cesair	'čuorpmas'	*ceseir-en*	'čuorpmasčalbmi'
Maltagiella	*xagħar*	'vuovttat'	*xagħr-a*	'vuoktačalbmi'
	ħaxix	'suoidni'	*ħaxix-a*	'suoidnečalbmi'
	silġ	'čuorpmas'	*silġ-a*	'čuorpmasčalbmi'

Vaikko čuovvovaš kapihttaliin vuojun sámegiela singulatiivvaide, de lea buorre fuobmát ahte sámegiella ii daninassii leat áidnalunddot, baicce dieđut kymri- ja maltagiela lágan gielain veahkehit min oaidnit sámegielas dakkár ieš vuođaid maidda eai gávdno čielga vástagat lagamus eanetlogugielain maid dutkantradišuvnnat leat leamaš vuođđun maiddái sámegiela giellaoahpaide.

3. Sámegielaid *-čalbmi* singulatiivva dovddaldahkan

Sámegielaid *-čalbmi* ii oro goassege guorahallon giellaoahpain dahje muđui giellaoahpa dutkama geahččanguovllus. Dát ii dattetge mearkkaš dan, ahte *čalbmi*-morfema singulatiivalágan geavahus ii livčče leamaš oahpis juo vuosttaš sámegiela dutkiide, geat eai namuhan fenomena iežaset giellaoahpain; sii dahke dan iežaset sátnegirjjiin:


ZHJALBME, et Øje, oculus. 2 en Partikel af en Samling, saa-ſom: et Sands-Korn, en Draabe Blod, particula compoſiti cujusdam, ut: granum arenæ, gutta ſangvinis. 3 en Maſke i et Garn, macula retis. pl: zhjalmek.

Saddo-zhjalbme (à ſaddo & zhjalb-me, qvod vide) et Sands-Korn, calculus arenæ.

Varra-zhjalbme (à var & zhjalb-me, qvod vide) en Blods-Draa-be, gutta ſangvinis.

Dás Knud Leem (1768) addá davvisámegiela *čalbmi*-sátnái golbma mearkkašumi: vuosttamužžan oaidninorgána ja goalmmádin mearkkašumi 'fierbmečalbmi', muhto dan ovdal nubbin mearkkašupmin lea 'en Partikel af en Samling', nugo dajaldagain *sáttočalbmi* 'et Sands-Korn' ja *varračalbmi* 'en Blods-Draabe'. Seamma áššii lei juo ovdal su fuomášuvvon máddelis, go Petrus (Pehr) Fiellström (1738: 144) muitalii ahte ruotagiela *sandkorn* lei sámegillii *saddetialme*. Dás ii leat vejolaš vuodjut buot sátnegirjjiide, muhto maiddái Nielsen (1932 s.v. *čâl'bme*) muitala, ahte dát mearkkašupmi lea 'single particle of something', ja su ovdamearkkat leat *káffečalbmi*, *sáttočalbmi*, *muohtačalbmi*, *arvečalbmi* ja *suoldnečalbmi*. Sammallahti (2021) ođđasamos sátnegirjjis dákkár sánit leat juo logiid mielde: omd. *s*-álgosaččat leat *sáltečalbmi*, *sáttočalbmi*, *siehppačalbmi*, *sitnočalbmi*, *sohkarčalbmi*, *suoidnečalbmi* ja *suoldnečalbmi*. Dattetge dát ovdanbuktojit ovttaskas sátnin mat gullet sátnegirjjiide muhto eai oro dán rádjai gullan giellaoahppagirjjiide.

Ieš goittotge jurddašan, ahte bajábeale dajaldagat gullet giellaohppii seamma ollu go sátnegirjjiide. Jus diehtit – maiddái mii geat leat oahppan sámegiela giellaoahppagirjjiin eatge eatnigiellan – ahte *-čalbmi* lea giellaoahpalaš elemeanta man sáhttá lasihit ávnnassániide, eat dárbbaš jur smiehttat, maid mearkkašit omd. *bihcečalbmi*, *borgačalbmi*, *hillačalbmi*, *muoldačalbmi* ja *nisočalbmi*, vaikko dat orrot váilumin buot sámegielaid sátnegirjjiin:[1]

[1] Buot dán artihkkala ovdamearkacealkagiid lean gávdnan – nu go láven – buhttemeahttun divrras SIKOR-korpusis, mii ii jáhku mielde oba gávdnoše sámegielaid oahppiid, dutkiid ja eará geaveaheaddjiid illun almmá Trond Trosterud barggu haga.

(1) *Lottánjávrri alde rámškkuha soajáidis čearret, mii šealggáhallá ja bilaida dego* ***bihcečalbmi****, mii nástin gearrá muohtagierragis.*
(2) *Lea fiertu, veaháš biegga ja veaháš* ***borgačalmmit****.*
(3) *Ánddar lea goahtedáhku nala jođđan šattažit; dalle eai* ***hillačalmmit*** *gul cahkket goikeseammaliid buollit.*
(4) *Ja du maŋisboahttit šaddet nu eatnagin go* ***muoldačalmmit*** *eatnama alde, – –*
(5) *1,25 lihtter čáhci, 50 g jeasta, 4 tb sálti, 1,5 dl biellemas siepmanat, 2 dl olles* ***nisočalmmit****, 3 dl roava nisojáffut – –*

Mii sáhttit dulkot dáid sániid riekta, go diehtit ahte *bihci*, *borga*, *hilla*, *muolda* ja *nisu* čujuhit dihtolágan ávdnasiidda, materiálaide, dahje dain leat juohke dáhpáhusas dakkár refereanttat maid eat dábálaččat rehkenastte: eai láve leat ”okta bihci”, ”guokte muoldda” dahje ”golbma nisu”. Sátnegirjjiin váilot maid *bihcečalbmi* ja *muorječalbmi* mat liikká áinnas geavahuvvojit, ja dalle maid lea sáhka das, ahte geavahit morfema *-čalbmi* čujuhan dihte dakkár entitehtii mii lea juo Leem (1768) sániiguin ’en Partikel af en Samling’. Muhto kymri-, malta- ja máŋgga eará giela giellaoahpa dutkiid sániiguin omd. *bihcečalbmi* sáhttá dulkojuvvot singulatiivan – giellaoahpa mearridan hápmin substantiivii. Dovddaldatkeahtes ovttaidlogu hápmi *bihci* ii dábálaččat čujut ovtta bihcái muhto baicce bihcái ávnnasin; duohta ovttaidlogu mearkkašumi *bihci* oažžu easka dalle, go *bihcečalbmi* dan muitala. Lea dattetge fuomášan veara, ahte vaikko *bihcečalbmi* (1) ja eará sullasaš sánit čujuhit ovttaskas partihkkaliidda, de dain sáhttá dasto ráhkadit máŋggaidlogu hámiid dego *borgačalmmit* (2), *hillačalmmit* (3), *muoldačalmmit* (4) ja *nisočalmmit* (5), mat čujuhit máŋgga ovttaskas entitehtii oktanaga – seamma láhkai go rehkenasttehahtti substantiivvaid máŋggaidlogu hámit dábálaččat.

Vaikko bajábeale ovdamearkkat bohtet davvisámegielas, de lea dehálaš fuobmát, ahte sullasaš singulatiivvat gávdnojit miehtá Sámi. Muhtun sámegielat leat ain oalle uhccán govviduvvon, muhto singulatiivvat eai oro gieđahallon ovttage sámegiela giellaoahpain makkárge namahusain (gč. dattetge Nielsen 1926: 306 gii ovtta linnjás namuhastá hámiid *vuovttat* ja *vuoktačalbmi*). Leksikála perspektiivvas dákkár ”čalmmiid” leat oanehaččat gieđahallan maid Aikio (2009: 61–63) gárjilgiela loatnasáni *čilmu* ’rihpa’ oktavuođas ja Helander (2016: 43) ges sámegiela skuvlaohppiid movttiidahttima guorahaladettiin.[2] Dan dihte lea vuogas čájehit eanet cealkkaovdamearkkaid eará sámegielain:

Anárašgiella
(6) *Ko muáttá stuorrâ* ***muotâčoolmijd****, te enâmist lii talle* ***hablâmuotâ*** *adai* ***muotâ*** *lii haablâs.*
(7) *– – mun jiem tieđe maht lii ko táán Njellim kuolbânist lâi ovdil* ***jäävvil*** *já tääl lii puoh kuorbâm nuuvt, et ij* ***jäävvilčalme*** *oinuu já tääbbin lâi muorâ já lijjii pikkâšeh, mut muoi Ristnáá-Piättár-Mattijn juuđijm ubbâ peeivi já tilá te finniim pikkâšijd, puoh lii poldum.*
(8) *Niŋálâs torske kođá miljovn* ***meiničalmed*** *50–150 meetter jieŋŋâlvuotân.*
(9) *Taan räi kavnum* ***kollečoolmijn*** *stuárráámus teedij 393 gr já tot kavnui Suáluičielgi kuávlust.*

Julevsámegiella
(10) *Dan diehti dát ájnna ålmåj guhti juo jábmema lahka lij, oattjoj máttov lågodibme degu alme náste ja merragátte* ***sáttojtjalme****.*
(11) *Valla, jus li edna* ***jiegŋatjalme****, tjásjgoajkátja jali stuorra* ***dubmetjalme*** *atmosferran, de tjuovgga almmemáddaga guovlluj máhttá árrot gebmusap ja vielggadabbon, – –*
(12) *Da lidjin gájka nåv gievra ja fámulattja, sijá buohta mån dåbddiv iehtjam virddujiddje* ***muohtatjalmmen****.*
(13) *Ja de diedij bájkev gånnå agev lidjin riek ållo sare, ja gå dåhku jåvsådij, de ájtsaj ij lim åvvånis* ***muorjjetjalmme*** *dåppe.*

Máttasámegiella

[2] Aikio (2012: 62–63) namuha maid, ahte sámegiela *šalbmi* (omd. *áibmešalbmi*, *ákšošalbmi*, *nállošalbmi*) lea suopmelaš loatnasátni, man vuođđun lea suomagiela *silmä* ’čalbmi’. Máŋga sámegielaga geavahit dattetge *šalbmi*-sáni vásttan sáni *čalbmi*.

(14) *Dïhte akteges jeahna **ebrietjelmide** ååredæjjese damta.*
(15) *Lasth jïh **lopme-tjelmieh** hispieh.*
(16) *Daelie jis maahta aereden fahkedh, biejjhguakoe, mohte jueskie guktie kraesine jih jeatjah sjædtojne **suelnietjelmieh** guhkiem aeredsbiejjiem doekoe.*
(17) *Saajve-Biehtere nåejtietjaetsiem boengeste ohtsede jih naan **tjaetsietjelmetjh** Saajve-Læjsan tjelmide beaja.*

Dát ovdamearkkat muitalit ollu sámegielaid *čalbmi*-singulatiivvaid birra. Okta dábálamos konteaksta lea dalle go hállat ovtta dahje máŋgga muohtačalmmi birra (*muotâčalme* (6), *muohtatjalmme* (12), *lopme-tjelmie* (15)), ja sullasaš meteorologalaš sánit leat *jiekŋa*, *arvi* ja *suoldni* (julevsámegiela *jiegŋatjalmme* (11), máttasámegiela *ebrietjelmie* (14) ja *suelnietjelmie* (16)). Daidda laktása maid čáhci sihke arvvis ja muđui golggossátnin. Vaikko čáhcečalbmi dahjege *tjaetsietjelmie* lea juo uhcci, de vel uhcibun daid govvidit diminutiivasuorgádusat *tjaetsietjelmetje* dahje *tjaetsietjalmetje* 'čáhcečalmmáš' (17). Eanet fásta partihkkaliidda gullet meađđenčalbmi (*meiničalmi* (8)), juo badjelis namuhuvvon sáttočalbmi (*sáttojtjalmme* (10)) ja muorječalbmi (*muorjjetjalmme* (13)). Vel uhcibut leat gavjačalmmit (*dubmetjalme* (11)) maid mihtidit mikromehteriid ja nanográmmaid mielde, muhto oalle stuoris lea 393-grámmasaš gollečalbmi (*kollečalme* (9)). Buot dánsullasaš sánit gávdnojit maiddái davvisámegielas, muhto ovdamearkkas (7) namuhuvvon *jäävvilčalme* ('jeagilčalbmi') lea sátni masa in leat fuobmán autenttalaš vástagiid eará gielain.

Lea maid fuomášan veara, ahte máŋgga ovdamearkacealkagis vuhtto čielga kontrásta singulatiivahámi ja dábálaš, ávnnassáni "ovttaidlogu" hámi gaskkas: Ovdamearkkas (6) daddjo, ahte anárašgillii gohčodit muohttaga (*muotâ*) namahusain *hablâmuotâ* dallego *muáttá stuorrâ muotâčoolmijd* dahjege muohttá stuorra muohtačalmmiid, ii omd. "stuorra muohttaga", man ávnnassániin livčči váttis dadjat. Seamma láhkai ovdamearkkas (7) lea sáhkan dat, ahte Njellima guolbanis gávdnui ovdal jeagil, muhto dál ii oidno ii jeagilčalbmige, ii okta bihtášge dan ávdnasis man ovdal vásihedje dego juogekeahtes ollisvuohtan. Ja de vel máttasámegiela *nåejtietjaetsie* (17) mii lea čáhci man sáhttá váldit ozas ja juohkit uhcit čáhcečalmmážiidda. Ovdamearkkas (15) ges sattáhallet ovttaskas lasttat (*lasth*) ja ovttaskas muohtačalmmit (*lopme-tjelmieh*), mii čájeha bures dan, ahte *laste* 'lasta' ii leat *lopme* lágan ávnnasátni ja danin dalle geavahit dábálaš máŋggaidlogu hámi eatge singulatiivvaid dego */??*lastetjelmie*(*h*).

Eará sámegielaid dego mat nuortalašgiela buohta eai gávdno nu stuorra dutkankorpusat go badjelis namuhuvvon gielaid buohta, muhto T. I. Itkonena (1958: 644–645) stuorra sátnegirji namuha omd. nuortalašgiela ovdamearkkaid *muõttčâ'lmm*, *čää'ccčâ'lmm*, *mue'rjjčâ'lmm*, *lue'mčâ'lmm* ja *sââ'rrčâ'lmm*. Aiddo dát sátnegirji čilge dáid singulatiivvaid mearkkašumi namalassii numerálain '1': *mue'rjjčâ'lmm* lea suomagillii '1 marja', *lue'mčâ'lmm* ges '1 hilla' ja *sââ'rrčâ'lmm* '1 mustikka'. Seamma gáldu sisttisdoallá lasseovdamearkkaid maiddái gieldda- ja darjjesámegielas (omd. 'jokŋa-', 'muohta-', 'varra-', 'sátto-' ja 'vuoktačalbmi').

Buohkanassii buot ovdamearkkat, maid dán rádjai lean ovdanbuktán, čájehit ahte *čalbmi*-elemeanttaid guovddáš funkšuvdna sámegielain lea earuhit ovtta dihto partihkkala dahje eanet ovttaskas partihkkaliid stuorát, daninassii juogekeahtes ollisvuođas, gavjačalmmážiid rájes gitta badjel miljovdna geardde stuorát ja miljárda geardde losit gollečalmmiid rádjai. Dákkár "čalmmiid" hámit varierejit sakka, mii orru čujuheamen dasa, ahte *čalbmi*-elemeanttas ii dábálaččat leat makkárge konkrehta leksikála mearkkašupmi, man hámi sáhtášii omd. sárgut.

Áidna stuorát erohus sámegielaid gaskka orru leamen dat, ahte máttasámegielas geavahit sáni *goelke-tjåelie* "guolgačoalli" dan sadjái mii davvin lea *guolgačalbmi*, ja maiddái julevsámegielas lea *vuobdda-tjalmme* lassin maid *vuobddatjoalle*, guktot jorgaluvvon sátnegirjjiin 'hårstrå'. Nuppe dáfus lea namuhan veara, ahte vaikko *vuodjačalbmi* ja dan vástagat eará davimus sámegielain eai oro geavahuvvomin neutrála singulatiivan muhto baicce seamma kulinára mearkkašumis go eanetlogugielaid *smørøye*, *smöröga* ja *voisilmä*, de máttasámegiela *voeje-tjelmie* lea sátnegirjji mielde dárogillii *fettprikk på vann* (Bergsland ja Magga 1993 s.v.), muhtunlágan singulára dat maid.

4. Digaštallan

Dán čállosa álggu rájes lean viggan čuoččuhit, ahte davvisámegiela *-čalbmi* ja dan vástagat eará sámegielain doibmet singulatiivva dovddaldahkan dalle, go elemeanta *-čalbmi* laktása substantiivii, man vuođđohápmi čujuha árgabeaieallimis uhcit eanet juogekeahtes ávdnasiidda dego *čáhci*, *muohta*, *varra*, *sáttu*, *suoidni* ja *vuovttat*. Dat boahtá ovdan earenoamáš konkrehtalaččat dalle go hállat omd. *muorječalmmi* ja *káffečalmmi* birra, dasgo daid lea veháš lunddolut rehkenastit go omd. *sáttočalmmiid* ja *suoldnečalmmiid*. Jus *čalbmi*-elemeantta oppalaš mearkkašumi galggašii čilget ovttain sániin, de dat lea lagamustá 'okta' dahje 'muhtun', ja aiddo dán tearpmain *singulatiiva* lávejit oaivvildit.

Dakkár funkšuvdna, mii ng. singulatiivahámiin láve gielain leat, gullá eanet giellaoahpa go sátnevuorkká beallái. Elemeanta *-čalbmi* lea dattetge olggosoaidnit dego substantiiva, mii laktása nuppi substantiivii, ja boađusin lea goallossátni. Dát leage árbevirolaš oaidnu das mo ja gos omd. *sáttočalbmi* ja *suoldnečalbmi* galget govviduvvot. Dán čállosis ii leat sadji vihkkedallat buot daid sivaid manin dát oaidnu maid lea áddehahtti ja manin in leat dan ollásit hilgumin. Háliidan baicce oanehaččat buohtastahttit morfema *-čalbmi* sajádaga kánske juohke sámegiela giellaoahpas dasa, maid lean eará sajiin čállán morfemaid *-ráigge* ja *-guovttos* birra.

Ylikoski (2014, 2015) guorahallá ng. prolatiiva sátnehámiid dego *bálggesrái*(*gge*), *bálgesrájge* ja *baalkaraejkiem*, mat orrot hámi dáfus dego čađačuovgi goallossánit, muhto daid funkšuvnnat sámegielain leat eanet kásusa láganat: *bálggesráigge* ii muital ráiggi birra – seamma láhkai go *suoldnečalbmi* ii muital dábálaš čalmmi birra. Dan sadjái dát hápmi sulastahttá earenoamážit báikekásusiid, maiddái syntávssa dáfus: *Mii? – Dat boares bálggis* : *Gos? – Da**n** boares bálgá**s*** : *Gokko? – Da**n** boares bálgges**rái**(**gge**)*. Aikio ja Ylikoski (2022: 157, 174) ges čilgeba, ahte vaikko morfema *-guovttos* (~ *-guoktá*) čielgasit laktása numerálii *guokte* ja orru sáni lágan – ja geavahuvvo sátnin goit gihpus *dat guovttos* – de dattetge dat geavahuvvo dábálaččat sátnehámiin dego *nieiddaguovttos* ja *váhnenguovttos*, main dan mearkkašupmi lea guvttiidlogu mearkkašupmi, vaikko vel geavahuvvoge dábálaččat dušše definihtta olmmošpáraid birra.[3] Dát guokte morfema eaba dieđus leat seamma čielga sojahangehčosat go lokatiivva *-s* dahje máŋggaidlogu *-t* – eaba seamma oanehaččat, opáhkat dahje seammá bákkolaš oasit sámegiela giellaoahpas. Dattetge dat leaba dan mađe giellaoahpalaš morfemat, ahte daid sáhttá buohtastahttit maiddái čielga sojahangehčosiiguin. Dán čállosa fáttá, *čalbmi*-singulatiivva, dilli lea maiddái aiddo dákkár: ii áibbas giellaoahpalaš, muhto ii suige áibbas leksikála morfema, dábálaš substantiiva. Buot dát "ođđa" sojahanmorfemat orrot leamen muhtun láhkai giellaoahpa ja sátnevuorkká rájá alde.

Badjelis lean karakteriseren *čalbmi*-hámiid álkivuođa dihte singulatiivan, muhto sámegielaid singulatiivvain lea dat ieš vuohta, ahte *čalbmi*-loahppasaš singulatiivvat leat čielgasit ovttaidlogu hámit, singulárat. Viehka dávjá dat dattetge geavahuvvojit máŋggaidlogu hámiin dego *suoldnečalmmit*, čujuhan dihte eai dušše suoldnái dihtolágan ávnnasin dahje (ovtta) suoldnečalbmái muhto baicce máŋgga suoldnečalbmái. Dát ii leat singulatiiva-doahpaga prototyhpalaš funkšuvdna, muhto dattetge dábálaš gielain, main lea singulatiivamorfemat mat leat čielgasit affivssat – ovdamearkka dihte kymrigielas (Haspelmath ja Karjus 2017: 1221; Grimm 2018: 529, 533–534). Dákkár singulatiivvaid lea vejolaš gohčodit dárkilut namahusain *individualiserejeaddji* (*individualizer*; gč. omd. Haspelmath 2019).

Loahpas sáhttá namuhit vel ovtta earenoamáš ieš vuođa, man maiddái sáhttá dulkot singulatiivvaid vuollešládjan. Michaelis (2013) hállá ng. antiduála birra, mainna son čujuha dasa mo muhtun gielain earenoamážit páralaš rumašlahtuid ovttaskas láhtuide sáhttá čujuhit earenoamáš vugiiguin dalle, go háliida deattuhit ahte hállá dušše ovtta oasi birra. Sámegielain oainnat geavahit elemeantta *-čalbmi* maiddái dalle:

(18) *Gili čađa váccedettiin Aili luittii gieđa Elvi ruoidna **oalgečalmmi** ala oadjudan dihte su.*
(19) *Su eallin gal ii rievdda, vaikko vel leage addán nuppi **manинčalmmi** vielljasis.*

[3] Hárvenaš muhto máŋgga eatnigielaga mielde giellaoahpalaš spiehkastagat oidnojit ovdamearkkain (i–ii), main definihtta guvttiidlogu hámit leat ráhkaduvvon sániin mat čujuhit heakkahis tiŋggaide. Guktot ovdamearkkat bohtet Gunvor Guttorm duodjái guoskevaš dieđalaš artihkkaliin, main duojár-čállis lea hui lagas oktavuohta gahpirguoktáin ja báhkiguoktáin maid analysere:

(i) *Muhto dán ge **gahperguoktás** leat stuorra erohusat.*
(ii) *Oainnán **báhkiguoktá** leaba dego jumežat.*

Sullasaš hámit leat maid *bálločalbmi* ja julevsámegiela *buolvvatjalmme* ’čibbedákti’. Dákko maid oaidnit, ahte *-čalbmi* addá rumašlahttonamahusaide mearkkašumi ’okta’ dahje ’muhtun’ dahje sáhttá máŋggaidlogu hámiin individualiseret daid refereanttaid eanet go dábálaččat (*oalgečalmmit*, *manінčalmmit*).

Dán čállosa álggus čujuhin dasa, ahte sámegiela *čalbmi* lea okta boarrásamos árbesániin, mat leat gávdnon juo urálalaš vuođđogielas. Dás ii leat vejolašvuohta vuodjut sámegielaid singulatiivvaid álgohistorjái, muhto oanehaččat sáhttá namuhit, ahte vuođđourálalaš **ćilmä* ’čalbmi’ maŋisboahttiin leat oalle seammalágan singulatiivva funkšuvnnat maid eará gielain, earenoamážit Sibirjjá manjsi-, hanti- ja samojedagielain, ja muhtun muddui maiddái uŋgáragielas. Dáid gielaid singulatiivvaide lea gidden eanet fuomášumi easka Däbritz (2021). Ieš lean eará sajis (Ylikoski 2021) evttohan, ahte muhtun sámegielaid singulatiivvaid sáhtášii Sibirjjá gielaid vuođul rekonstrueret gitta urálalaš vuođđogillii. Dákkár sánit sáhtášedje leat omd. **mẹrja-ćilmä* ’muorječalbmi’ ja **lumi-ćilmä* ’muohtačalbmi’ dahjege máttasámegillii *lopmetjelmie*, ja earenoamážit **weri-ćilmä* ’varračalbmi’ ja **jäŋi-ćilmä* ’jiekŋačalbmi’, maidda orrot gávdnomin dárkilis vástagat nuge gáiddus gielain go davvihantigielas (*wŭr sem* ’varračalbmi’) ja uŋgáragielas (*jégszem* ’čuorpmasčalbmi’). Maiddái julevsámegiela *buolvvatjalmme* ’čibbedákti’ sáhtášii adnojuvvot árbin hámis **pu/oxli-ćilmä*, mii lea vuovdeenetsagielas otnon hápmin *fuase* ’čibbi’ (Aikio 2012: 230) ja selkupagielas seilon goallossátnin *пулсаŭ* (*pulsaj*) ~ *пулхаŭ* (*pulxaj*) (Bykonja et al. 2005 s.v.). Lea maid miellagiddevaš fuobmát, ahte dánlágan singulatiivvat gávdnojit aiddo dain urálalaš giellabearraša gielain, main gávdno maiddái guvttiidlohku. Sámegielaid lassin Eurohpá bealde eai leat dákkár urálalaš gielat; muđui Eurohpá urálalaš gielain numerusvuogádagat leat seammaláganat go ovdamearkka dihte suomagielas: ovttaidlogut ja máŋggaidlogut, muhto guvttiidloguid ja singulatiivvaid haga.

Dán oktavuođas lea maid vuogas namuhit, ahte sullasaš singulatiivvat, maid vuođđun leat substantiivvat mearkkašumiin ’čalbmi’, gávdnojit maiddái muhtun eará gielain, muhto dábálaččat dat eai gal leat. Urálalaš giellabearraša nuorttimus rávddas Gaska-Sibirjjás hállojuvvo kettagiella, mii lea jeniseilaš giellabearraša maŋimuš ealli giella. Doppe maid leat singulatiivvat dego mat *eːlʲ* ’muorji’ → *eːlʲdes’* ’muorječalbmi’, *qoː* ’čuorpmas’ → *qoːdes’* ’čuorpmasčalbmi’ ja *hʌn’aŋ* ’sáttu’ → *hundes’* ~ *hʌ́n’aŋdis’* ’sáttočalbmi’, ja dán singulatiivagehčosa *-des’* álgovuođđun atnet substantiivva **des* ’čalbmi’, mii lea seilon sátnin *hās’* ’čalbmi’ (Helimski 2016: 157–159). Eanet čielga parallealat eai dattetge oro álkit gávdnamis, muhto liikká lea vejolaš gávdnat sullasaš fenomenaid maiddái omd. Afrihkás: Mátta-Sudana luwogielas čujuhit vuonccesmoađi (*jén*) ovttaskas loddái sániin *wóŋ jén*, bustávalaččat ”čalbmi vuonccis” dahjege várra ”vuonccesčalbmi” (Storch 2014: 278–279).

5. Loahpahus

Dán čállosis lean geahččan, dulkon ja čilgen sámegiela *čalbmi*-morfema singulatiiva geavahanvugiid dihtomielalaččat giellaoahpa perspektiivvas. Fertet dattetge muitit, ahte dát lea vuosttaš dutkamuš mii geahččala áddet *suoldnečalmmiid*, *suoidnečalmmiid* ja máŋga eará *-čalmmi* dán ođđa geahččanguovllus, ja máŋga detálja báhcet ain guorahalakeahttá. Dárkilut guorahallama haga báhcá earret eará dat, ahte muhtun suopmaniin *luomečalbmi* čujuha ovtta olles muorjái, muhtun eará suopmaniin ges dakkár ovttaskas – na, luomečalbmái – mat leat ovtta muorjjis máŋga, jus ii leat ng. ovttačalmmat luomi. Namuhan veara leat maid ovdamearkka (3) *hillačalmmit*, mat ”eai – – cahkket goikeseammaliid buollit”. Dát ovdamearka boahtá davvisámegiela oarjjimus suopmaniin ja soaitá orrut amas máŋgga guovllus, muhto dása maid gávdno miellagiddevaš parallealla selkupagielas, man sáni *tü haj* ’čuonan’ etymologalaš vástta sámegillii livčči *”dollačalbmi”* (< urálalaš vuođđogiela **tuli* + **ćilmä*). Lassedutkamušat áinnas dárbbašuvvojit, ja okta dutkanfáddá livčče maiddái erohusat dábálaš máŋggaidlogu hámiid ja singulatiivva máŋggaidlogu hámiid semantihkas (omd. *guolggat* vs. *guolgačalmmit*, *vuovttat* vs. *vuoktačalmmit* muhto maid omd. *muohttagat* vs. *muohtačalmmit*).

Vaikko lean dán čállosis háliidan čuoččuhit, ahte dás guorahallon sámegielaid *čalbmi*-dajaldagaid sáhtášii atnit dihtolágan singulatiivahápmin, de dat eai suige leat seamma čielga sojahanhámit go giela dábálamos sojahanhámit (dahje omd. kymri- ja maltagiela singulatiivvat oanehis gehčosiiguin *-yn*/*-en* ja *-a*; gč. kapihttala 2). *Čalbmi*-singulatiivvaid sáhttá baicce buohtastahttit maiddái *ráigge*-prolatiivvain ja *guovttos*-duálain, mat leat sámegiela goallossániid ja sojahanhámiid rájá alde – nubbi (ja dávjá historjjálaš) juolgi sátnevuorkkás, nubbi ges nannosit giellaoahpa bealde.

Gáldut

Aikio, Ante. 2009. *The Saami loanwords in Finnish and Karelian*. Nákkosgirji, Oulu universitehta, Oulu.

Aikio 2012 = Luobbal Sámmol Sámmol Ánte (Aikio, Ante). 2012. On Finnic long vowels, Samoyed vowel sequences and Proto-Uralic *x. Girjjis *Per Urales ad Orientem: Iter polyphonicum multilingue. Festskrift tillägnad Juha Janhunen på hans sextioårsdag den 12 februari 2012*, doaimmahan Tiina Hyytiäinen, Lotta Jalava, Janne Saarikivi ja Erika Sandman, s. 227–250. Société Finno-Ougrienne, Helsinki.

Aikio, Ante (Luobbal Sámmol Sámmol Ánte) ja Jussi Ylikoski. 2022. North Saami. Girjjis *The Oxford guide to the Uralic languages*, doaimmahan Marianne Bakró-Nagy, Johanna Laakso ja Elena Skribnik, s. 147–177. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198767664.003.0010.

Bergsland, Knut ja Lajla Mattsson Magga. 1993. *Åarjelsaemien daaroen baakoegærja. Sydsamisk norsk ordbok*. Idut.

Bykonja, Valentina, Nadežda Kuznecova ja Natal′ja Maksimova. 2005. *Sel'kupsko-russkij dialektnyj slovar'*. Izdatel′stvo Tomskogo gosudarstvennogo pedagogičeskogo universiteta, Tomsk.

Dali, Myriam ja Eric Mathieu. 2021. Singulative systems. Girjjis *The Oxford handbook of grammatical number*, doaimmahan Patricia Cabredo Hofherr ja Jenny Doetjes, s. 275–290. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780198795858.013.13.

Däbritz, Chris Lasse. 2021. Typology of number systems in languages of Western and Central Siberia. *Finnisch-Ugrische Forschungen* 66: 85–138. https://doi.org/10.33339/fuf.97288.

Fiellström, Petrus. 1738. *Dictionarium Sueco-Lapponicum*. Stockholm.

Grimm, Scott. 2018. Grammatical number and the scale of individuation. *Language* 94(3): 527–574. https://doi.org/10.1353/lan.2018.0035.

Haspelmath, Martin. 2019. A discussion with Edith Moravcsik about singulative markers and individualizers. Diversity Linguistics Comment. Posted on 2019-06-26 by Martin Haspelmath. https://dlc.hypotheses.org/1808.

Haspelmath, Martin ja Andres Karjus. 2017. Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency. *Linguistics* 55(6): 1213–1235. https://doi.org/10.1515/ling-2017-0026.

Helander, Nils Øivind. 2016. *Ohppojuvvon ja sohppojuvvon giella. Gielladiđolašvuohta, čálamáhttu ja guovttegielatvuohta*. Sámi allaskuvla, Guovdageaidnu.

Helimski, Eugene. 2016. *S*-singulatives in Ket. *Journal of Language Relationship* 14(3): 157–163. https://doi.org/10.31826/jlr-2017-143-404.

Itkonen, T. I. 1958. *Koltan- ja kuolanlapin sanakirja*. Suomalais-Ugrilainen Seura, Helsinki.

Leem, Knud. 1768. *Lexicon Lapponicum bipartitum. Pars prima. Lapponico - Danico - Latina*. Nidaros.

Michaelis, Susanne Maria. 2013. Antidual of paired body-part terms. Duojis *The atlas of pidgin and creole language structures*, doaimmahan Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath ja Magnus Huber. Max Planck Institute for Evolutionary Anthropology, Leipzig. https://apics-online.info/parameters/27.chapter.html.

Nielsen, Konrad. 1926. *Lærebok i lappisk. I. Grammatikk*. A.W. Brøgger, Oslo.

Nielsen, Konrad. 1932. *Lappisk ordbok. Grunnet på dialektene i Polmak, Karasjok og Kautokeino. Vol. 1. A–F*. Aschehoug, Oslo.

Sammallahti, Pekka. 2021. *Sámi-suoma sátnegirji. Pohjoissaame – suomi -sanakirja*. Davvi Girji, Kárášjohka.

SIKOR = SIKOR. UiT Norgga árktalaš universitehta ja Norgga Sámedikki sámi teakstačoakkáldat. Veršuvdna 01.10.2021. http://gtweb.uit.no/korp/.

Storch, Anne. 2014. *A grammar of Luwo. An anthropological approach*. Benjamins, Amsterdam. https://doi.org/10.1075/clu.12.

Ylikoski, Jussi. 2014. Davvisámegiela *-ráigge* – substantiiva, advearba, postposišuvdna vai kásus? *Sámi dieđalaš áigečála* 2/2014: 47–70.

Ylikoski, Jussi. 2015. From compounds to case marking: Prolatives in South Saami and Lule Saami. *Finnisch-Ugrische Mitteilungen* 39: 101–155.

Ylikoski, Jussi. 2021. Ice eyes, blood eyes: Remarks on the Uralic singulative marker **ćilmä* ‘eye’. Språkets funktion 17, 26.–27.5.2021, Åbo. https://www.academia.edu/49054434/Ice_eyes_blood_eyes_Remarks_on_the_Uralic_singulative_marker_ćilmä_eye.
Zeuss, J. C. 1853. *Grammatica Celtica*. Weidmann, Lipsia.

Trond Trosterud – publikasjonar 1989–2022

Her er ei liste over publikasjonane til Trond Trosterud frå den første i 1989. Sjøl om lista er lang, er ho nok ikkje heilt komplett.

Trosterud, Trond. 1989. The null subject parameter and the new mainland Scandinavian word order – a possible counterexample from a Norwegian dialect. In *Papers from the Eleventh Scandinavian Conference of Linguistics*, edited by Jussi Niemi. Vol. 1: 87–100. Joensuu.

Trosterud, Trond and Sjur Nørstebø Moshagen. 1990. Non-Clause-Bounded Reflexives in Mainland Scandinavian. *Working Papers in Scandinavian Syntax* 46: 47–52.

Trosterud, Trond. 1991. Sidontasuhteita Ruijansuomessa. In *Papers from the eighteenth Finnish Conference of Linguistics*, pp. 54–66.

Trosterud, Trond. 1991. Ruijansuomen sidontasuhteita. In *Nordlyd, Tromsø University Working Papers on Language & Linguistics* 20: 86–97.

Trosterud, Trond. 1991. Lokalkasus og preposisjonar i finsk, kvensk, samisk og norsk. *Norsk lingvistisk tidsskrift* Årgang 9: 50–78.

Trosterud, Trond. 1992. Binding relations in two Finnmark Finnish dialects. A comparative syntactic study. Hovudoppgåve i lingvistikk 1990. *Working papers in linguistics* 12. University of Trondheim.

Trosterud, Trond. 1993. Anaphors and Binding Domains in Finnish. *Case and Other Functional Categories in Finnish Syntax*. De Gruyter Mouton, pp. 225–243. https://doi.org/10.1515/9783110902600.225.

Trosterud, Trond. 1993. Estiske V2-tendensar. I *Flyktförsök. Kalasbok till Christer Platzack på femtioårsdagen 18 november 1993, från doktorander och dylika*. Lunds universitet, Lund.

Trond Trosterud, Anders Holmberg, Urpo Nikanne, Irmeli Oraviita and Hannu Reime. 1993. The Structure of INFL and the Finite Clause in Finnish. In *Case and other Functional Categories in Finnish*. De Gruyter Mouton, pp. 177–206. https://doi.org/10.1515/9783110902600.177.

Trosterud, Trond. 1994. Auxiliaries, negative verbs and word order in the Sami and Finnic languages. In *Minor Uralic languages: Structure and development*, edited by Ago Künnap. Tartu, pp. 173–181. http://hdl.handle.net/10062/54646.

Trosterud, Trond. 1994. Miksi mordvan objektikonjugaatio näyttää juuri sellaselta kuin se näytää? *Volgalaiskielet muutoksessa*. Volgalaikielten symposiumi Turussa 1–2.9.1993. Turun ylipoiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja 45: 74–86. https://finna.fi/Record/arto.014411014.

Trosterud, Trond. 1995. The objective conjugation in the Uralic languages – a unified approach. In *Congressus octavus internationalis Fenno-Ugristarum, Jyväskylä 10.–15.8.1995 Pars 2 : Summaria acroasium in sectionibus et symposiis factarums*, pp. 119–120.

Trosterud, Trond. 1995. Sovjetisk språkpolitikk. Finsk-ugriske språk i Russland i eit historisk-politisk perspektiv. *Nordisk Østforum* 3 – 1995: 40–45.

Trosterud, Trond. 1995. V2 og V3 i estisk. *Norsk lingvistisk tidsskrift* Årgang 13: 187–198.

Trosterud, Trond. 1996. Den finske dialektkrigen (1820-1840). I *Ivar Aasen-konferansen til Norsk Målungdom. Mål og makt* 1–2/1996: 130–151.

© 2022 *Nordlyd* 46.1: 309–316, *Morfologi, målstrev og maskinar: Trond Trosterud {fyller | täyttää | deavdá | turns} 60!*, redigert av Lene Antonsen, Sjur Nørstebø Moshagen og Øystein A. Vangsnes. Publisert ved UiT Noregs arktiske universitet. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.6671

Trosterud, Trond. 1996. Funny characters on the net. How information technology may (or may not) do support minority languages. *Arbete Människa Miljö & Nordisk Ergonomi* 1996.

Trosterud, Trond. 1996. Die südsamische Wortfolge als eine Kombination der deutschen und marischen Wortfolgen analysiert. Im Lars Gunnar Larsson (Hrsg.): *Lapponica et Uralica. 100 Jahre finnisch-ugrischer Unterricht an der Universität Uppsala. Studia Uralica Upsaliensia* 26: 103–112. Uppsala.

Trosterud, Trond. 1997. On supporting threatened languages. *Iatiku, Newsletter of the Foundation for Endangered Languages* 4: 22–24.

Trosterud, Trond. 1997. Humanistic Research in Arctic Russia. *Sustainable Development in the Arctic. Consequences of Industrial Development in the European Arctic Regions. Report from an interdisciplinary network conference an workshop, Trondheim, Norway, 5.–7. February 1997*. CED-Report 2/97, pp. 98–102.

Trosterud, Trond. 1997. Kveeni kirjakielenä? *Ruijan kaiku – Norjan suomalainen lehti* 1997:10, s. 9.

Trosterud, Trond. 1998. Finsk på fem minutter. *Ruijan kaiku – Norjan suomalainen lehti* 1998:5, s. 15.

Trosterud, Trond. 1998. To teser om norsk målstrid. *Norsk Tidend* 3/98: 4–5.

Trosterud, Trond. 1998. Vi har plutseleg fått to ordbøker mellom estisk og norsk. *Ord om ord* 4: 77–79. Årsskrift for leksikografi, Universitetet i Oslo, Institutt for nordistikk og litteraturvitenskap, Oslo.

Trosterud, Trond. 1998. Den første estisk-norske ordlista er publisert. Turid Farbregd / Gennadi Jagomägi Eesti-norra sõnastik / Estisk-norsk ordliste. *LexicoNordica* 5: 275–280. https://tidsskrift.dk/lexn/article/view/18871.

Trosterud, Trond. 1998. Yksi suomi ja neljä skandinaaviskaa – vai päinvastoin? (Summary: One Finnish and four Scandinavian written languages – or the other way round?). *Nordlyd, Tromsø University Working Papers on Language & Linguistics* 26: 27–35.

Trosterud, Trond. 1998. Lillehammerske og aasenske nasjonalismar – Nasjonalisme. *Mål og makt* 28(1): 4–7.

Trosterud, Trond. 1999. Formen skal være Een – men hvilken? Om nynorsknormalen. *Mål og makt* 29(1): 28–30.

Trosterud, Trond. 1999. Review. Pekka Sammallahti: The Saami Languages. An Introduction. *Nordic Journal of Linguistics*, Vol. 22(1): 91–94. https://doi.org/10.1080/03325869950137072.

Trosterud, Trond. 1999. (Bokmelding) Turid Farbregd, Sigrid Kangur og Ülle Viks: Norra-eesti-Eesti-norra sõnaraamat /Norsk-estisk– Estisk-norsk ordbok. *LexicoNordica* 6: 241–248. https://tidsskrift.dk/lexn/article/view/18990.

Trosterud, Trond. 1999. Koding av namn i folkeregisteret sine databasar. *Nytt om namn* 30: 13–18.

Trosterud, Trond. 2000. [Bokmelding av] Torbjørn Nordgård (red.): Innføring i språkvitenskap. *Norsk Lingvistisk Tidsskrift* 18: 264–278.

Trosterud, Trond. 2000. Festtalen. *Ruijan kaiku. Norjan suomalainen lehti* 2000:6, s. 15.

Trosterud, Trond. 2001. [Bokmelding av] Kåven, Brita E. (red) 2000: Stor norsk-samisk ordbok / Dárusámi sátnegirji. *LexicoNordica* 8: 283–306. https://tidsskrift.dk/lexn/article/view/18764.

Hagen, Kristin, Pia Lane og Trond Trosterud. 2001. En grammatikkontroll for bokmål. *Språknytt* 3: 6–9.

Trosterud, Trond. 2001. The changes in Scandinavian morphology from 1100 to 1500. *Arkiv för nordisk filologi* 116: 153–191.

Trosterud, Trond. 2001. Genustilordning i norsk er regelstyrt. *Norsk lingvistisk tidsskrift* Årgang 19: 29–58.

Trosterud, Trond. 2002. Morfologiija rolla sámi giellateknologiijas. *Sámi dieđalaš áigečála* 1/2002: 90–105.

Trosterud, Trond. 2002. Ut i verda: sidemålsopplæring som norsk eksportartikkel. *Globalisering og språkpolitikk*. Noregs Mållag, Oslo, s. 30–40.

Trosterud, Trond. 2002. Parallel corpora as tools for investigating and developing minority languages. In *Parallel corpora, Parallel worlds. Language and Computers*, edited by Lars Borin. Studies in practical linguistics no 43. Rodopi, Amsterdam, pp. 111–122. https://doi.org/10.1163/9789004334298_007.

Trosterud, Trond. 2003. Språkdaude, purisme og språkleg revitalisering. I *Purt og reint. Om purisme i dei nordiske språka*, redigert av Brunstad, Brodersen og Sandøy. Vol. Nr 15, Skrifter frå Ivar Aasen-instituttet. Høgskulen i Volda, Volda, s. 181–216. https://www.academia.edu/30721381/Spr%C3%A5kdaude_purisme_og_spr%C3%A5kleg_revitalisering.

Trond Trosterud. 2003. [Bokmelding av] Språkteigen / Sylfest Lomheim. *Språknytt* 31(1/2): 34–35, 40.

Trosterud, Trond. 2003. Ordbokskritikk. *LexicoNordica* 10: 65–88. https://tidsskrift.dk/lexn/article/view/19786.

Trosterud, Trond. 2003. Morfologi og leksikalsk semantikk. Patrik Bye, Trond Trosterud og Øystein Vangsnes: *Språk og språkvitskap. Ei innføring i lingvistikk*. Det Norske Samlaget, Oslo, s. 49–122.

Trosterud, Trond. 2004. Porting morphological analysis and disambiguation to new languages. *First Steps in Language Documentation for Minority Languages: Proceedings of the SALTMIL Workshop at LREC 2004*. European Language Resources Association, Paris, pp. 90–92.

Moshagen, Sjur og Trond Trosterud. 2005. Samisk språkteknologi. I *Nordisk sprogteknologi. Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000–2004*. Museum Tusculanums Forlag, Københavns universitet, s. 57–60.

Trosterud, Trond. 2005. Samisk og kvensk i Noreg etter 1905. *Språknytt* 2005(1–2): 43–47. https://www.sprakradet.no/Vi-og-vart/Publikasjoner/Spraaknytt/Arkivet/2005/Spraaknytt_1-2_2005/Samisk_og_kvensk/.

Nesset, Tore og Trond Trosterud. 2005. Ny norsk-russisk ordbok: Ei leksikografisk storhending. *LexicoNordica* 12: 273–284. https://tidsskrift.dk/lexn/article/view/18671.

Trosterud, Trond and Heli Uibo. 2005. Consonant Gradation in Estonian and Sami: Two-Level Solution. In *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday*, edited by Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund and Anssi Yli-Jyrä. CSLI Studies in Computational Linguistics online, pp. 136–150. https://web.stanford.edu/group/cslipublications/cslipublications/koskenniemi-festschrift/14-trosterud-uibo.pdf.

Moshagen, Sjur, Pekka Sammallahti and Trond Trosterud. 2005. Twol at work. In *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday*, edited by Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund and Anssi Yli-Jyrä. CSLI Studies in Computational Linguistics online, pp. 94–105. https://web.stanford.edu/group/cslipublications/cslipublications/koskenniemi-festschrift/10-moshagen-sammallahti-trosterud.pdf.

Trosterud, Trond. 2006. Gender assignment in Old Norse. *Lingua* Vol 116/9: 1441–1463. https://doi.org/10.1016/j.lingua.2004.06.015.

Gaup, Børre, Sjur Nørstebø Moshagen, Thomas Omma, Maaren Palismaa, Tomi Pieski and Trond Trosterud. 2006. From Xerox to Aspell: A First Prototype of a North Sámi Speller Based on TWOL Technology. In *Finite-State Methods and Natural Language Processing. Lecture Notes in Computer Science* 4002, edited by Anssi Yli-Jyrä, Lauri Karttunen and J. Karhumäki. Springer-Verlag, Berlin – Heidelberg, pp. 306–307. https://doi.org/10.1007/11780885_37.

Trosterud, Trond. 2006. Grammar-based Language Technology for the Sámi Languages. In *Lesser used Languages & Computer Linguistics.* Europäische Akademie, Bozen, pp. 133–148. https://doi.org/10.1515/9783110197785.

Trosterud, Trond. 2006. Grammatically based language technology for minority languages. In *Lesser-Known Languages of South Asia. Status and Policies, Case Studies and Applications of Information Technology*. Mouton de Gruyter, Berlin, pp. 293–316. https://doi.org/10.1515/9783110197785.

Trosterud, Trond. 2006. *Homonymy in the Uralic Two-Argument Agreement Paradigms*. Mémoires de la Société Finno-Ougrienne 251. https://www.sgr.fi/fi/items/show/109.

Trosterud, Trond, Saara Huhmarniemi and Sjur Nørstebø Moshagen. 2007. Usage of XSL Stylesheets for the Annotation of the Sami Language Corpora. In *LAW '07: Proceedings of the Linguistic Annotation Workshop*. Association for Computational Linguistics, Morristown, NJ, USA, pp. 45–48. https://doi.org/10.3115/1642059.1642066.

Trosterud, Trond og Linda Wiechetek. 2007. Disambiguering av homonymi i nord- og lulesamisk. I *Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007*. Suomalais-Ugrilaisen Seuran Toimituksia 253. Suomalais-Ugrilainen Seura, Helsinki, s. 401–421. https://www.sgr.fi/sust/sust253/sust253_trosterudjawiechetek.pdf.

Trosterud, Trond. 2008. [Bokmelding] Verbh. En sydsamisk verbhandbok. *LexicoNordica* 15: 347–354. https://tidsskrift.dk/lexn/article/view/18519/16192.

Muhirwe, Jackson and Trond Trosterud. 2008. Finite State Solutions For Reduplication In Kinyarwanda Language. *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pp. 73–80. https://aclanthology.org/I08-3013/.

Trosterud, Trond. 2008. Language Assimilation During the Modernisation Process: Experiences from Norway and North-West Russia. *Acta Borealia* 25(2): 93–112. https://doi.org/10.1080/08003830802496653.

Trond Trosterud. 2008. [Bokmelding] Viveca Rabb, Genuskongruens på reträtt. Variation i nominalfrasen i Kvevlaxdialekten. *Svenska landsmål och svenskt folkliv* 2008: 175–177.

Moshagen, Sjur Nørstebø och Trond Trosterud. 2008. Datorstöd för samiska och andra minoritetsspråk. I *Tekniken bakom språket*, redigerad av Rickard Domeij. Småskrift utarbetad av språkrådet. Norstedts akademiska Förlag, Falun.

Tyers, Francis M., Linda Wiechetek and Trond Trosterud. 2009. Developing Prototypes for Machine Translation between Two Sámi Languages. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation, EAMT09*. European Association for Machine Translation, Allschwil. https://aclanthology.org/2009.eamt-1.17/.

Unhammer, Kevin and Trond Trosterud. 2009. Reuse of Free Resources in Machine Translation between Nynorsk and Bokmål. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, edited by J.A. Pérez-Ortiz, F. Sánchez-Martínez and F.M. Tyers. Alacant, Spain, pp. 35–42. http://rua.ua.es/dspace/handle/10045/12025.

Trosterud, Trond. 2009. A constraint grammar for Faroese. *NEALT Proceedings Series*, Vol 8: 1–7. https://dspace.ut.ee/bitstream/handle/10062/14289/proceedings.pdf.

Antonsen, Lene, Trond Trosterud, Ciprian-Virgil Gerstenberger og Sjur Nørstebø Moshagen. 2009. Ei intelligent ordbok for samisk. *LexicoNordica* 16: 271–283. https://tidsskrift.dk/lexn/article/view/18479.

Antonsen, Lene, Berit Anne Bals Baal, Saara Huhmarniemi ja Trond Trosterud. 2009. Dihtor ja giela válljenvejolašvuođat – gielalaš ja pedagogalaš čuolmmat. (The Computer and the Variability of Language – Linguistic and Pedagogical Issues). Girjjis Johanna Ijäs ja Nils Øivind Helander (doaim.), *Sáhkavuoruin sáhkan. Sámegiela ja sámi girjjálašvuođa muhtin áigeguovdilis dutkanfáttát*. Dieđut 1/2009. Sámi allaskuvla, Guovdageaidnu, s. 87–102. https://giellatekno.uit.no/publications/Diedut_2009_1_samlet_20091015_pdfX3-2002.pdf.

Antonsen, Lene, Saara Huhmarniemi and Trond Trosterud. 2009. Constraint Grammar in Dialogue Systems. *NEALT Proceedings Series 2009*. Volum 8: 13–21. https://dspace.ut.ee/bitstream/handle/10062/14287/antonsen_etal.pdf.

Antonsen, Lene, Saara Huhmarniemi and Trond Trosterud. 2009. Interactive pedagogical programs based on constraint grammar. In *Proceedings of the 17th Nordic Conference of Computational Linguistics.* Nealt Proceedings Series 4, pp. 10–17. https://dspace.ut.ee/handle/10062/9546.

Trosterud, Trond. 2010. Felles leksikalske ressursar for språkteknologi og leksikografi. *LexicoNordica* 17: 211–223. https://tidsskrift.dk/lexn/article/view/18632/16288.

Antonsen, Lene ja Trond Trosterud. 2010. Manne dihtor galgá máhttit grammatihka? [English summary: Why the computer should know its Sami grammar.] *Sámi dieđalaš áigečála* 1/2010: 3–28. http://site.uit.no/aigecala/sda-1-2010-antonsen-ja-trosterud/.

Antonsen, Lene, Linda Wiechetek and Trond Trosterud. 2010. Reusing Grammatical Resources for New Languages. In *Proceedings of the International conference on Language Resources and Evaluation LREC 2010*. The Association for Computational Linguistics, Stroudsburg, pp. 2782–2789. http://www.lrec-conf.org/proceedings/lrec2010/pdf/254_Paper.pdf.

Antonsen, Lene and Trond Trosterud. 2011. Next to nothing – a cheap South Saami disambiguator. *NEALT Proceedings Series 2011*. Volum 14 [10]: 1–7. https://dspace.ut.ee/bitstream/handle/10062/19296/antonsen_trosterud.pdf.

Trosterud, Trond og Hilde Skanke. 2012. Kvensk juridisk terminologi. *Terminologen* Volum 1: 56–63.

Тростерюд, Тронд. 2012. Роль языковой технологии в сохранении и ревитализации языка. In Саамская идентичность: проблемы сохранения языка и культуры на Севере: Материалы международной научной конференции. Murmansk State Humanities University 2012, 3–11.

Trosterud, Trond. 2012. A restricted freedom of choice: Linguistic diversity in the digital landscape. *Nordlyd, Tromsø University Working Papers on Language & Linguistics* 39(2): 89–104. https://doi.org/10.7557/12.2474.

Trosterud, Trond and Berit Merete Nystad. 2012. A North Sami translator's mailing list seen as a key to minority language lexicography. In *Euralex 2012 Proceedings: Euralex International Association for Lexicography* 2012, pp. 250–256. https://www.euralex.org/elx_proceedings/Euralex2012/pp250-256%20Trosterud%20and%20Eskonsipo.pdf.

Johnson, Ryan, Lene Antonsen and Trond Trosterud. 2013. Using finite state transducers for making efficient reading comprehension dictionaries. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), May 22–24, 2013*, Oslo University, Norway. NEALT Proceedings Series 16: 59–71. https://aclanthology.org/W13-5610.pdf.

Trosterud, Trond and Kevin Brubeck Unhammer. 2013. Evaluating North Sámi to Norwegian assimilation RBMT. In *Free/Open-Source Rule-Based Machine Translation*, edited by Cristina España-Bonet and Aarne Ranta, pp. 13–26. http://www.grammaticalframework.org/~aarne/FreeRBMT-2012.pdf.

Moshagen, Sjur Nørstebø, Tommi Pirinen and Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa 2013), May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series* 16: 343–352. https://aclanthology.org/W13-5631.pdf.

Antonsen, Lene, Ryan Johnson, Heli Uibo and Trond Trosterud. 2013. Generating modular grammar exercises with finite-state transducers. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NoDaLiDa 2013, May 22–24, Oslo, Norway. NEALT Proceedings Series* 17: 27–38. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=86&Article_No=3.

Moshagen, Sjur, Jack Rueter, Tommi Pirinen, Trond Trosterud and Francis M. Tyers. 2014. Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages. *Workshop: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*. LREC 2014, pp. 71–77. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014-Workshop-CCURL2014-Proceedings.pdf.

Haavisto, Mervi, Kaisa Maliniemi, Leena Niiranen, Pirjo Paavalniemi, Tove Reibo og Trond Trosterud. 2014. Kvensk ordbok på nett – hvem har nytte av den? I *Nordiske Studier i Leksikografi* (NSL). https://tidsskrift.dk/nsil/article/view/20997.

Snoek, Conor, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Nørstebø Moshagen and Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 34–42. https://doi.org/10.3115/v1/W14-2205.

Sjaggo, Ann-Charlotte og Trond Trosterud. 2015. Om pitesamiskt språk. *Från kust til kyst = Áhpegáttest áhpegáddáj*, reidert av Bjørg Evjen og Marit Myrvoll. Orkana Forlag, Tromsø, s. 223–231.

Trosterud, Trond. 2015. Grønlandsk, samiske språk og den nordiske språkdeklarasjonen. *Sprog i Norden*, 2015: 131–140. https://tidsskrift.dk/sin/issue/view/3333.

Trosterud, Trond ja Marja-Liisa Olthuis. 2015. Inarinsaamen lingvistinen suunnittelu kieliteknologian valossa. *Agon – Pohjonen Tiede- ja kulttuurilehti* 1–2/2015. http://agon.fi/article/inarinsaamen-lingvistinen-suunnittelu-kieliteknologian-valossa/

Antonsen, Lene, Trond Trosterud and Francis Tyers. 2016. A North Saami to South Saami Machine Translation Prototype. *Northern European Journal of Language Technology (NEJLT). NEJLT* vol. 4. https://doi.org/10.3384/nejlt.2000-1533.1642.

Harrigan, Atticus G., Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology* 27(4): 565–598. https://doi.org/10.1007/s11525-017-9315-x.

Arppe, Antti, Jordan Lachler, Trond Trosterud, Lene Antonsen and Sjur Nørstebø Moshagen. 2016. Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. In *Proceedings of the LREC 2016 Workshop "CCURL 2016 – Towards an Alliance for Digital Language Diversity"*, pp. 1–8. https://altlab.ualberta.ca/wp-content/uploads/2017/05/LREC_CCURL_Arppe_et_al_2016C.pdf.

Trosterud, Trond. 2017. Cafe Boddu lei mu sámegiel orrunlatnja. *Sámis, Sámi čálakultuvrralaš áigečála* 25: 58–59. https://www.samifaga.org/samis/samis25//files/assets/common/downloads/publication.pdf

Trosterud, Sindre Reino, Trond Trosterud, Anna-Kaisa Räisänen, Leena Niiranen, Mervi Haavisto and Kaisa Maliniemi. 2017. A morphological analyser for Kven. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages W17–0608*, pp. 76–88. https://doi.org/10.18653/v1/W17-0608.

Antonsen, Lene og Trond Trosterud. 2017. Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk. *Norsk lingvistisk tidsskrift* 35(2): 153–185. https://hdl.handle.net/10037/13250.

Johnson, Ryan, Tommi Pirinen, Tiina Puolakainen, Francis Morton Tyers, Trond Trosterud and Kevin Unhammer. 2017. North Sámi to Finnish rule-based machine translation system. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017, Gothenburg, Sweden*. NEALT Proceedings Series. Linköping University Electronic Press, Linköpings universitet. pp. 115–122. https://ep.liu.se/ecp/131/014/ecp17131014.pdf.

Antonsen, Lene, Ciprian-Virgil Gerstenberger, Maja Lisa Kappfjell, Sandra Nystø Rahka, Marja-Liisa Olthuis, Trond Trosterud and Francis Morton Tyers. 2017. Machine translation with North Saami as a pivot language. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017, Gothenburg, Sweden*. Number 131 in Linköping Electronic Conference Proceedings. Linköping University Electronic Press, Linköpings universitet. pp. 123–131. https://ep.liu.se/ecp/131/015/ecp17131015.pdf.

Morottaja, Petter, Marja-Liisa Olthuis, Trond Trosterud já Lene Antonsen. 2018. Anarâškielâ tivvoomohjelm – Kielâ- já ortografiafeeilâi kuorrâm tivvoomohjelmáin. *Dutkansearvvi dieđalaš áigečála* 2(2): s. 63–84. https://www.dutkansearvi.fi/volume-2-issue-2-en/.

Domeij, Rickard, Ola Karlsson, Sjur Nørstebø Moshagen and Trosterud, Trond. 2019. Enhancing Information Accessibility and Digital Literacy for Minorities Using Language Technology—the Example of Sámi and Other National Minority Languages in Sweden. In *Perspectives on Indigenous writing and literacies.* Brill Academic Publishers 2019, pp. 113–137. https://doi.org/10.1163/9789004298507_007.

Kaalep, Heiki-Jaan, Sjur Nørstebø Moshagen and Trond Trosterud. 2018. Estonian Morphology in the Giella Infrastructure. In *Human Language Technologies – The Baltic Perspective*. IOS Press 2018, pp. 47–54. https://ebooks.iospress.nl/volumearticle/50303.

Trosterud, Trond. 2019. Kva bruker vi minoritetsspråksordbøker til? Ein studie av brukarloggane for tolv tospråklege ordbøker. *LexicoNordica* 26: 177–202. https://hdl.handle.net/10037/18357.

Moshagen, Sjur Nørstebø, Trond Trosterud and Lene Antonsen. 2019. Language Technology for Indigenous Languages: Achievements and Challenges. In *Proceedings of the Language Technologies for All (LT4All). Paris: UNESCO Headquarters, 5–6 December 2019*, pp. 219–222. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.55.pdf.

Moshagen, Sjur Nørstebø and Trond Trosterud. 2019. Rich Morphology, No Corpus — And We Still Made It. The Sámi Experience". In *Proceedings of the Language Technologies for All (LT4All). Paris: UNESCO Headquarters, 5–6 December 2019*, pp. 379–383. https://lt4all.elra.info/proceedings/lt4all2019/pdf/2019.lt4all-1.95.pdf.

Trosterud, Trond og Lene Antonsen. 2020. Hva er viktig for forståelse? Om maskinoversetting fra nordsamisk. I K. Hagen, A. Hjelde, K. Stjernholm og Ø. A. Vangsnes (red.) *Bauta: Janne Bondi Johannessen in memoriam. Oslo Studies in Language*, 11(2): 489–502. University of Oslo, Oslo. https://doi.org/10.5617/osla.8514.

Antonsen, Lene og Trond Trosterud. 2020. Med et tastetrykk. Bruk av digitale ressurser for samiske språk. I *Samiske tall forteller* 13. Kommentert samisk statistikk 2020. Sámi allaskuvla. s. 1–26. https://samilogutmuitalit.no/se/node/4105.

Trosterud, Trond. 2020. Samiske bokutgjevingar i Noreg – eit uttrykk for norsk samepolitikk? I *Samiske tall forteller* 13. Kommentert samisk statistikk 2020. Sami allaskuvla. s. 85–102. https://samilogutmuitalit.no/se/node/4108.

Antonsen, Lene, Trond Trosterud, Linda Wiechetek og Chiara Argese. 2021. *Da forskere ved UiT skapte grunnlaget for en digital samisk skriftkultur*. Kronikk i forbindelse med Hjernekraftprisen 2021. Forskerforbundet. https://www.forskerforbundet.no/PageFiles/29257/Hjernekraftprisen2021_Antonsen-mfl.pdf.

Wiechetek, Linda, Chiara Argese, Tommi Pirinen and Trond Trosterud. 2021. Suoidne-varra-bleahkka-mála-bihkka-senet-dielku 'hay-blood-ink-paint-tar-mustard-stain' -Should compounds be lexicalized in NLP? In Johanna Monti, Felice Dell'Orletta and Fabio Tamburini (red.): *CLiC-it 2020, Proceedings of the Seventh Italian Conference on Computational Linguistics. Bologna, Italy, March 1–3, 2021*. https://doi.org/10.4000/books.aaccademia.8979.

Trosterud, Trond and Sjur Nørstebø Moshagen. 2021. Soft on errors? The correcting mechanism of a Skolt Sami speller. In *Multilingual Facilitation*, edited Hämäläinen, Partanen and Alnajjar, pp. 197−207. https://doi.org/10.31885/9789515150257.

Olthuis, Marja-Liisa, Trond Trosterud, Erika Katjaana Sarivaara, Petter Morottaja and Eljas Niskanen. 2021. Strengthening the Literacy of an Indigenous Language Community: Methodological Implications of the Project Čyeti čälled anarâškielân, 'One Hundred Writers for Aanaar Saami'. In *Indigenous Research Methodologies in Sámi and Global Contexts. New Research – New Voices*, Volume: 11. Brill, Leiden, pp. 175–200. https://brill.com/view/title/56605.

Trosterud, Trond. 2022. Utan tastatur, ingen tekst: om det språkteknologiske grunnlaget for språka våre. I *Framgång för små språk. En översikt om varför små språk i Norden behöver stärkas och vad som bidrar till ett lyckat språkstärkande arbete. Innehåller en checklista med framgångsfaktorer*, redigerad av Karin Kvarfordt Niia, s. 68–73. https://www.diva-portal.org/smash/get/diva2:1639815/FULLTEXT01.pdf.

Dominczak, Katarzyna, Lene Antonsen og Trond Trosterud. (Under produksjon). Fra partikkelverb og preposisjoner til verbavledninger og kasus. Brukerstudie av ei nordsamisk-norsk-nordsamisk ordbok. *LexicoNordica* 29.