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.

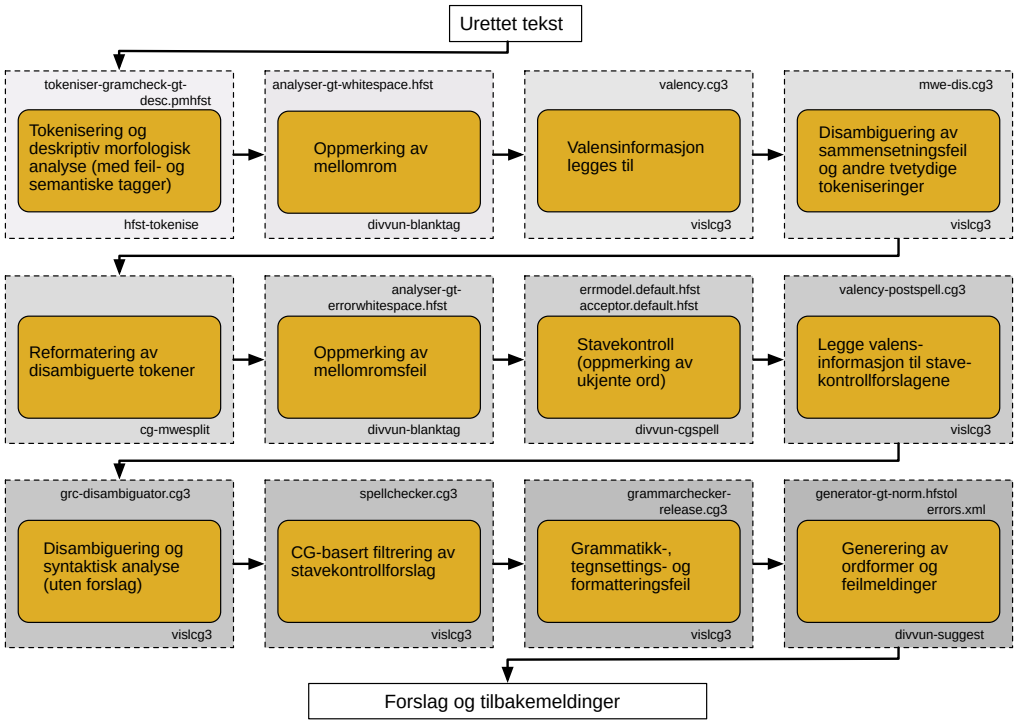
4. Metodevalg

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

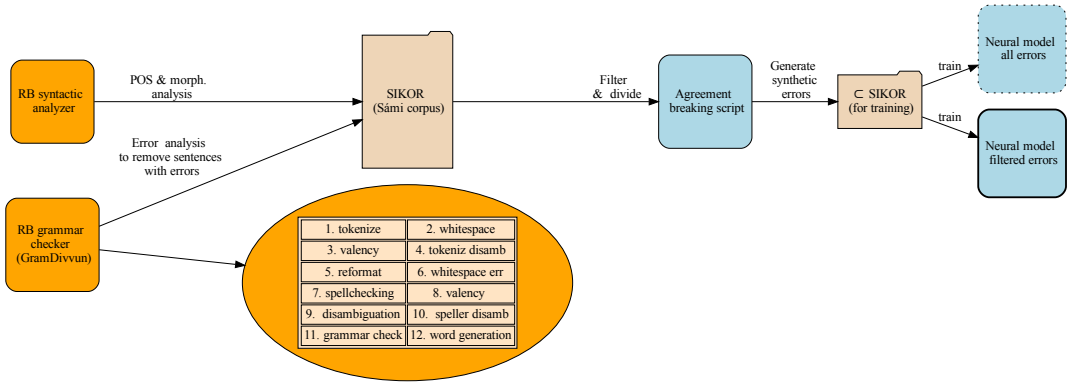
[18] Vi inkluderer hele konfigurasjonen av opennmt-py og skript til trening i https://gtsvn.uit.no/hybrid_gramcheck ved publisering

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.