LIA-korpusa – eldre talemålsopptak for norsk og samisk gjort tilgjengelege

Kristin Hagen[1] & Øystein A. Vangsnes[2,3]
[1]*Universitetet i Oslo,* [2]*UiT Noregs arktiske universitet,* [3]*Høgskulen på Vestlandet*

Abstract

This paper presents the results from the project *Language Infrastructure made Accessible* (LIA) which had as its main goal to digitize and make accessible old recordings of spoken Norwegian and Sámi from various archives, first and foremost from the four partner institutions University of Oslo, University of Bergen, The Norwegian University of Science and Technology, and UiT The Arctic University of Norway. The infrastructures resulting from the project can be summarized as 1) various language technology resources such as a morphological tagger and a parser for Norwegian dialects, upgrading of the corpus interface Glossa and a new infrastructure for file depots, 2) a file depot of Norwegian dialect recordings, 3) three corpora of spoken Norwegian and one for North Sámi as well as the LIA treebank. The paper exemplifies how the corpora can be utilized.

Keywords: corpus, spoken language, Norwegian dialects, North Sámi dialects, language infrastructure

1. Innleiing

Prosjektet *Language Infrastructure made Accessible* (LIA) var eit infrastrukturprosjekt i perioden 2014–2019 finansiert av Noregs forskingsråd gjennom deira infrastrukturprogram. Prosjektet vart gjennomført som eit samarbeid mellom Universitetet i Oslo, Universitetet i Bergen, Noregs teknisk-naturvitskaplege universitet, UiT Noregs arktiske universitet, Nasjonalbiblioteket og Norsk ordbok 2014. Initiativtakar til, og leiar av, prosjektet var professor Janne Bondi Johannessen (1960–2020) ved Tekstlaboratoriet, UiO.

Målet med LIA-prosjektet var for det første å digitalisere og gjere tilgjengeleg eldre opptak med norske og samiske dialekter som fanst ved universiteta, og for det andre å transkribere og annotere eit utval opptak av god kvalitet. Prosjektet har resultert i følgjande ressursar (sjå eiga liste med URLar til slutt i artikkelen):

1. Språkteknologiske ressursar som ein morfologisk taggar og ein parser for norske dialektar, oppgradering av søkjeprogrammet Glossa og nyutvikling av fildepot-infrastruktur.
2. Eit fildepot med norske dialektopptak
3. Fire søkbare talespråkskorpus og ein trebank i søkegrensesnittet Glossa:
   a) *CANS - amerikanordisk talespråkskorpus*
   b) *LIA Sápmi - Sámegiela hállangiellakorpus*
   c) *TAUS - Talemålsundersøkelsen i Oslo (B-serien)*
   d) *LIA norsk - korpus av eldre dialektopptak*
   e) *LIA-trebanken (også nedlastbar for språkteknologiske formål)*

Ut over desse ressursane vart det i 2021 publisert ei bok, *Språk i arkiva: Ny forsking om eldre talemål frå LIA-prosjektet*, redigert av Kristin Hagen, Gjert Kristoffersen, Øystein A. Vangsnes og Tor A. Åfarli. Boka, som er elektronisk fritt tilgjengeleg på http://omp.novus.no/index.php/novus/catalog/book/19, inneheld 13 artiklar av 25 ulike forfattarar basert på føredrag haldne på avslutningskonferansen for prosjektet i november 2019. Artiklane tar alle utgangspunkt i data henta frå eit av korpusa nemnde over og tener slik som døme på korleis ressursane kan brukast.

Fildepotet og korpusa er tilgjengelege med innlogging via Feide, CLARIN eller som lokalt godkjend brukar. Transkripsjonar og trebank er publiserte på CC-lisens (Creative Commons). I denne artikkelen skal vi skildre desse ressursane nærmare, og vi skal også vise korleis ein kan bruke dei. Størst vekt vil bli lagt på dei to korpusa for høvesvis samisk og norsk som begge ber prosjektakronymet i namna sine. Men før vi vender oss til desse ressursane, skal vi gi ei kort framstilling av historikken knytt til prosjektet.

 Nordlyd 47.2: 119–130, *Struktur, ideologi og mangfald*, redigert av Ragni Vik Johnsen, Carola Kleemann, Øystein A. Vangsnes & Maud Westendorp. Publisert ved UiT – Noregs arktiske universitet. http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.7157

2. LIA-prosjektet

I løpet av dei siste 60 åra har det vore samla inn mange talemålsopptak for ulike formål rundt omkring i Noreg. Nokre av dei har vore digitaliserte og katalogiserte på systematisk vis, andre har lege i arkivskåp og skuffar. Mange av dei har stått i fare for å bli øydelagde. LIA-prosjektet gjorde det mogleg å samle inn store mengder av desse opptaka frå dei fire norske universiteta som har eller har hatt eigne talemålsarkiv: Universitetet i Oslo, Universitetet i Bergen, Noregs teknisk-naturvitskaplege universitet og UiT Noregs arktiske universitet. Opptaka vart digitaliserte ved Nasjonalbiblioteket i Mo i Rana og kopiar er langtidslagra der, i alt over 3500 norske og nordsamiske lydfiler frå spoleband og kassettar.

Etter at filene var digitaliserte, vart dei handsama av ein av dei mange prosjekttilsette i prosjektet. Om lag 75 personar var involverte i prosjektet på ein eller annan måte, alt frå professorane i styringsgruppa, ingeniørane som organiserte og programmerte filarkiv og søkjegrensenitt, til deltidstilsette studentar som transkriberte, annoterte, korrekturlas og organiserte metadata. Prosjektet hadde deltidstilsette ved alle dei fire universiteta medan den daglege koordineringa skjedde frå Tekstlaboratoriet i Oslo.

Innhaldet i dei digitaliserte filene vart kartlagt og merka med mest mogleg metadata. Dei mest interessante opptaka med god kvalitet vart transkriberte. Dette var opptak med fri tale, ofte om tema som handverkstradisjonar, landbruk, matlaging og draktskikkar. Filer med til dømes opplesing eller oppramsing av bøyingsparadigme og stadnamn vart ikkje prioriterte. Utplukkinga av interessante filer vart gjort i samråd med styringsgruppa for prosjektet.

Dei samiske transkripsjonane vart ortografisk transkriberte. Dei norske fekk både ein talemålsnær variant og ein ortografisk nynorsk variant ved at dei talemålsnære transkripsjonane vart halvautomatisk translitterert og deretter korrigerte og korrekturlesne. Denne metoden vart utvikla for Nordisk dialektkorpus (NDK) i prosjektet Nordisk dialektsyntaks (Johannessen et al. 2009). Men medan den ortografiske transkripsjonen av den norske delen av NDK er til bokmål, vart det bestemt at den i LIA-korpuset skulle vere til nynorsk. Det kravde noko utviklingsarbeid som vi straks kjem tilbake til. Parallelt med dette arbeidet vart eit utval av transkripsjonane morfologisk og syntaktisk annoterte slik at vi kunne bygge opp ein trebank og trene ein morfologisk taggar og ein parser for nynorsk talemål.

Personvernombodet (SIKT, tidlegare NSD) har godkjent prosjektet og peiker på at personvernulempa ser ut til å vere liten og at samfunnsnytta overstig ulempa. Mange av talarane i korpuset lever ikkje lenger, sensitiv informasjon er tatt ut av korpuset og namn er anonymiserte. Korpus og fildepot er dessutan beskytta med innlogging.

Før vi gjer nærmare greie for korpusa og trebanken, vil vi kort summere opp arbeidet med den språkteknologiske delen av prosjektet.

3. Morfologisk taggar, parser og søkjeprogram for korpusa,

Gode og nyttige talespråkskorpus bør vere morfologisk og syntaktisk annoterte og ha eit enkelt og tilgjengeleg søkjeprogram. Sidan det ville bli særs kostbart å annotere alle transkripsjonane manuelt, valde vi å nytte automatiske verktøy sjølv om dei gjer ein del feil. For nynorsk fanst det frå før verken morfologisk taggar eller syntaktisk parser (automatiske verktøy for automatisk annotasjon av morfologisk og syntaktisk informasjon), så dette måtte vi utvikle i prosjektet. Vi starta med å annotere transkripsjonar frå 17 ulike stader i Noreg med Oslo-Bergen-taggaren for nynorsk (Johannessen et al. 2012). Sidan denne taggaren er utvikla for skriftspråk, måtte vi gå igjennom og rette opp resultatet før vi trente ein eigen statistisk taggar for nynorsk talespråk: LIA-taggaren. Den nyutvikla taggaren er målt til ei ordklassenøyaktigheit på 97,25 % ved ei tifaldig kryssvalidering. Lemmatisatoren, det vil seie analysatoren som knyter ordformer til rette oppslagsord, har ei nøyaktigheit på 96,88 %.

Parseren er laga på same måte. Først vart dei 17 transkripsjonane analyserte med ein dependensparser utvikla for skriftspråk, og deretter vart annoteringa manuelt retta opp. Resultatet vart LIA-trebanken (Øvrelid et al. 2018) som både er søkbar i Glossa og tilgjengeleg på Github i nedlastbar versjon. Vi fortel meir om Glossa-versjonen nedanfor. LIA-trebanken vart til slutt brukt for å trene ein dependensparser for nynorsk talemål.

LIA norsk er tagga med LIA-taggaren. LIA-korpusa CANS og TAUS er ortografisk transkriberte på bokmål og er også tagga med NoTa-taggaren som Tekstlaboratoriet har utvikla tidlegare (Nøklestad og

Søfteland 2007). På sikt vil alle dei norske LIA-korpusa få syntaktisk annotasjon med LIA-parseren og ein ny bokmålsparser for talespråk som er utvikla i Clarino+-prosjektet (Kåsen et al. 2022).

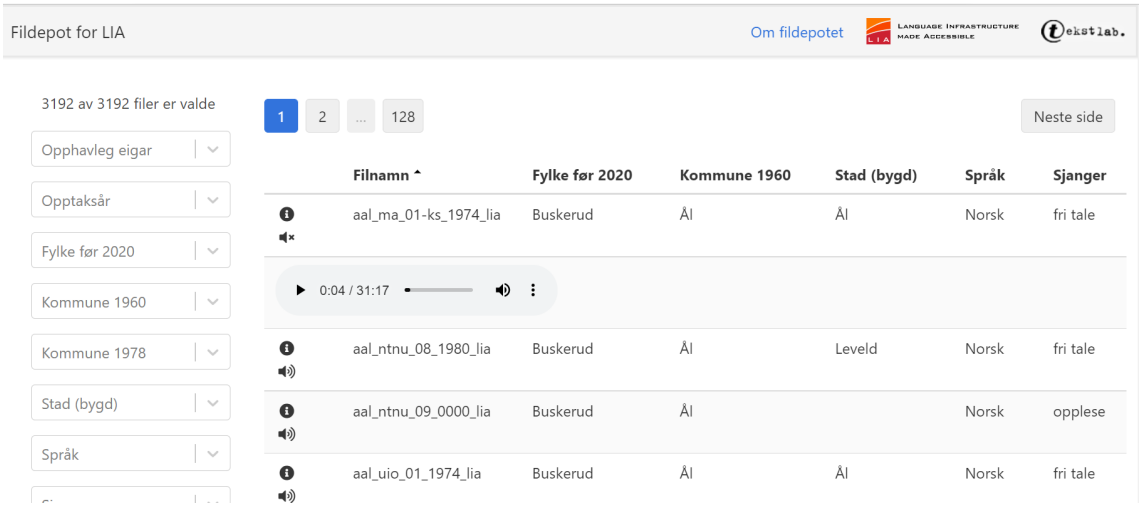
For samisk hadde LIA-prosjektet eit godt samarbeid med Giellatekno, forskingsgruppa for samisk språkteknologi ved UiT, som annoterte dei nordsamiske transkripsjonane med verktøy dei har utvikla i andre samanhengar.

Tekstlaboratoriet har gjennom mange år utvikla og videreutvikla søkjeverktøyet Glossa. Gjennom LIA-prosjektet og infrastrukturprosjektet Clarino+ er det laga ein ny versjon av Glossa som alle LIA-korpusa ligg i. Den nye versjonen av Glossa gjer det blant anna mogleg å søkje via kart, vise korpuset eller eit utval av korpuset i teksthandsamingsverktøyet Voyant og vise resultat av søk distribuert på metadata. Ein kan også gjere enkle syntaktiske søk og få resultata opp som dependensstrukturar. Alle transkripsjonane er tidskoda slik at ein kan høyre resultata av søk i korpuset. Vi vil vise fleire eksempel på søk og resultat–handtering i Glossa i kapittel 7 og 8.

4. LIA-fildepotet

Fildepotet frå LIA-prosjektet inneheld lydfiler, transkripsjonar og metadata frå den norske delen av prosjektet. Programmet for søk i fildepotet er nyutvikla i LIA-prosjektet, og nesten alle dei norske opptaka som vart digitaliserte, ligg i depotet. Unntaket er nokre få filer som er klausulerte. Fildepotet omfattar også dei filene som ikkje vart transkriberte. Til saman er det snakk om 3192 lydfiler der om lag 1300 har ein nedlastbar transkripsjon. Lydfilene er søkbare på metadata, og det er mogleg å spele dei av, men ikkje laste dei ned.

Det eldste opptaket i depotet er frå 1937, medan dei fleste er frå 1950-talet og fram til 1996. Sidan lydfilene i arkivet kjem frå fire universitet og er produserte av ulike prosjekt og forskarar til ulike tider, har det vore ei utfordring å finne og standardisere metadata. Nokre filer har hatt mykje informasjon, andre nesten ingenting. I prosjektet har vi forsøkt å gi alle filer informasjon om fylke, kommunenamn i 1960, kommunenamn i 1978, språk, sjanger og i mange tilfelle også stad (bygd). Alt dette er søkbart. Ved å trykke på informasjonsknappen ved sidan av filnamnet, er det mogleg å sjå all informasjon om opptaket.



Figur 1: Utsnitt av fildepotet for LIA. Den søkbare metadatamenyen er til venstre. Eit klikk på informasjonsknappen til venstre for filnamnet gir all metadatainformasjon om fila. Ved å klikke på lydsymbolet, kan ein høyre på fila og spole fram og tilbake.

Dei samiske opptaka er førebels ikkje inkluderte i fildepotet. Grunnen er at det er knytt større krav til personvern ved opptak med etniske minoritetar. I fildepotet kan ein høyre heile lydfiler, og få med seg heile historier og forklaringar. I det samiske talespråkskorpuset LIA Sápmi er talarane og historiene deira betre beskytta ved at det berre er mogleg å høyre korte utdrag i samband med lingvistiske søk.

5. CANS - amerikanordisk talespråkskorpus v. 3.1

Første versjonen av CANS – amerikanordisk talespråkskorpus vart lansert allereie i 2012 gjennom prosjektet NorAmDiaSyn. Seinare har korpuset vorte utvida fleire gonger gjennom ulike prosjekt og med ulik finansiering, sist gjennom LIA-prosjektet. No i 2023 er korpuset lagt inn i den nyaste versjonen av Glossa og inneheld nesten 775 000 ord og skiljeteikn (kalla «token» i grensesnitta). Det er flest norsk-amerikanarar i korpuset, men også nokre informantar med svensk opphav/bakgrunn. Nesten alle informantane har norsk eller svensk som morsmål, men mange har snakka mest engelsk etter at dei vart vaksne. Transkripsjonane finst i to versjonar: ein talemålsnær og ein ortografisk bokmål eller svensk.

Dei fleste opptaka i korpuset er gjort mellom 2010 og 2016 av Janne Bondi Johannessen og hennar kollegaer. Dei svenske opptaka er gjort av Ida Larsson med fleire mellom 2011 og 2014. CANS inneheld også eldre materiale frå Didrik Arup Seip og Ernst W. Selmer (1931), Einar Haugen (1942) og Arnstein Hjelde (1987, 1990, 1992).

Prosjektet *Norwegian across the Americas* ved Universitetet i Bergen (2020–2024) undersøker norsk språk over heile det amerikanske kontinentet og korleis det har utvikla seg over fleire generasjonar. Gjennom dette prosjektet har CANS fått transkribert meir av det eldre materialet, og i 2024 vil korpuset få nye opptak av talarar med norsk herkomst busett i Latin-Amerika der spansk er hovudspråket.

6. TAUS - Talemålsundersøkelsen i Oslo

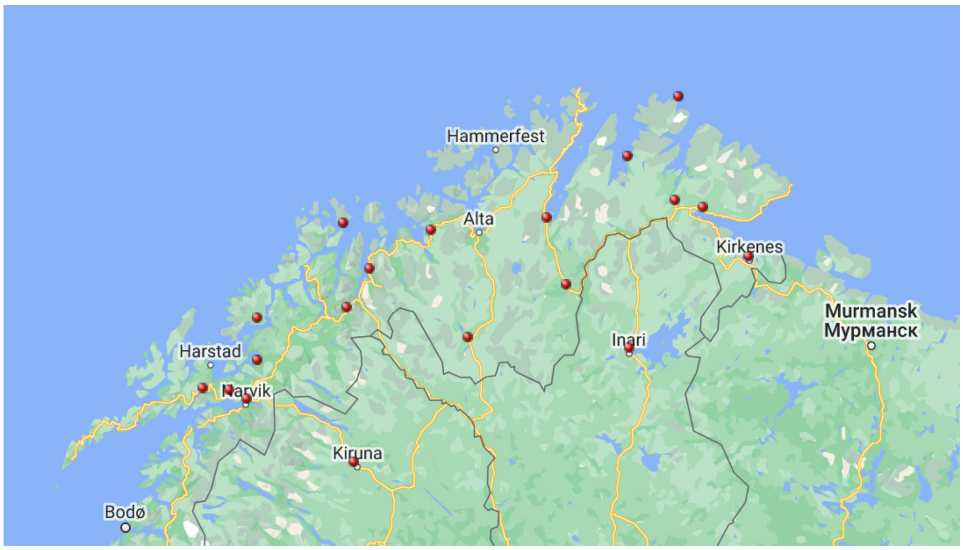
TAUS-korpuset er eit av dei eldste talespråkskorpusa til Tekstlaboratoriet. Opptaka er frå byrjinga av 1970-åra, og siktemålet for prosjektet den gongen var å granske sosiale skilnader i Oslo-målet. A og C-serien av TAUS vart nytranskribert og lagt i korpus på midten av 2000-talet. På det tidspunktet var opptaka i B-serien borte, og då dei vart funne att i eit bortgøymt arkiv i 2013, fanst det inga finansiering til å transkribere desse intervjua og samtalane frå Oslo. Gjennom LIA-prosjektet fekk lydbanda nytt liv, og TAUS-korpuset vart komplett.

TAUS-korpuset inneheld no nesten 388 000 ord og skiljeteikn med 86 informantar frå Oslo aust og vest. Korpuset er ortografisk transkribert til bokmål, men er kopla saman med dei originale TAUS-transkripsjonane der desse var moglege å finne fram til.

7. LIA Sápmi - Sámegiela hállangiellakorpus

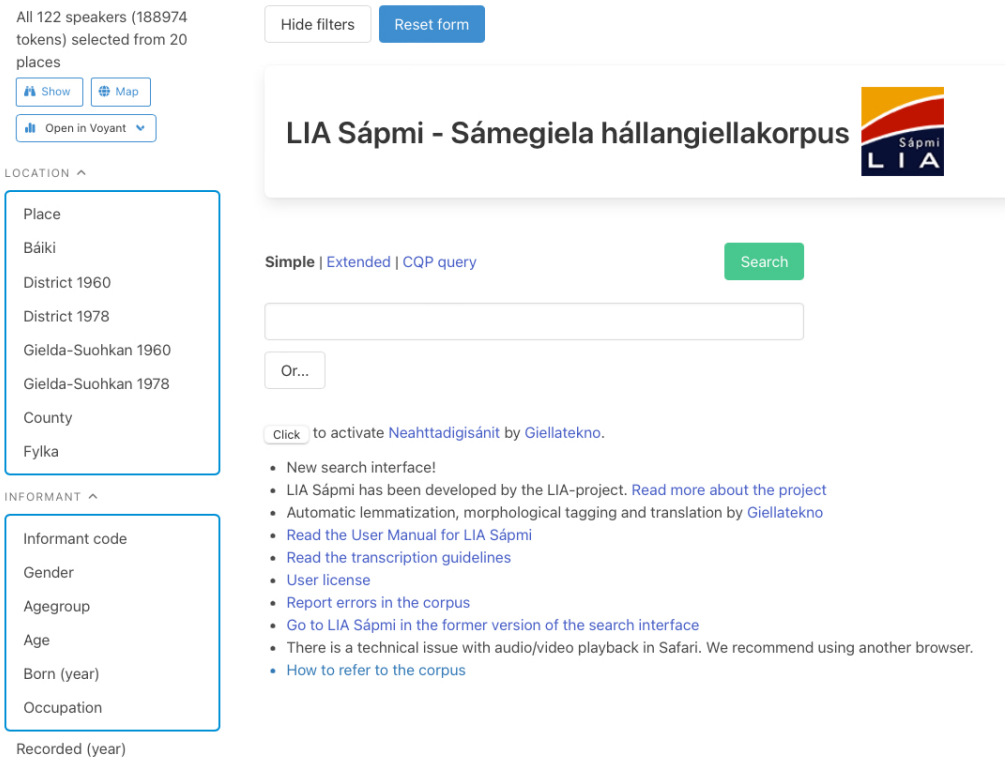
LIA Sápmi er det første talespråkskorpuset for nordsamiske dialektar, og det har komme til gjennom tett samarbeid med Giellatekno og tilsette på samisk ved UiT Noregs arktiske universitet. Sidan det samiske språksamfunnet er lite, viste det seg vanskeleg å få tak i kvalifiserte deltidstilsette som kunne transkribere og korrekturlese transkripsjonane. Ein valde difor å konsentrere arbeidet om ortografisk transkripsjon for å få transkribert og tilrettelagt flest moglege opptak.

Opptaka i LIA Sápmi er frå 1960 til 1987, og mange av dei kjem frå samlinga til Nils Jernsletten, tidlegare professor i samisk språk ved UiT. Då Jernsletten gjekk av med pensjon etterlét han seg ei stor øskje med band og kassettar med opptak som han dels hadde gjort sjølv av ulike, særleg nordsamiske, dialektar og dels kopiar av opptak frå andre kjelder, mellom anna NRK. I noko tid vart desse banda brukte i undervisning og forsking av kollegaer av Jernsletten, og i samband med LIA-prosjektet nytta ein høvet til å få dei digitaliserte og sikra på ein god måte. Desse opptaka utgjer grunnlaget for LIA Sápmi, eit korpus som inneheld nesten 190 000 ord og skiljeteikn frå 122 informantar som kjem frå 19 stader i det nordlege Sápmi, fremst frå norsk side, men også frå svensk og finsk side. Kart 1 viser kva stader korpuset har opptak frå.



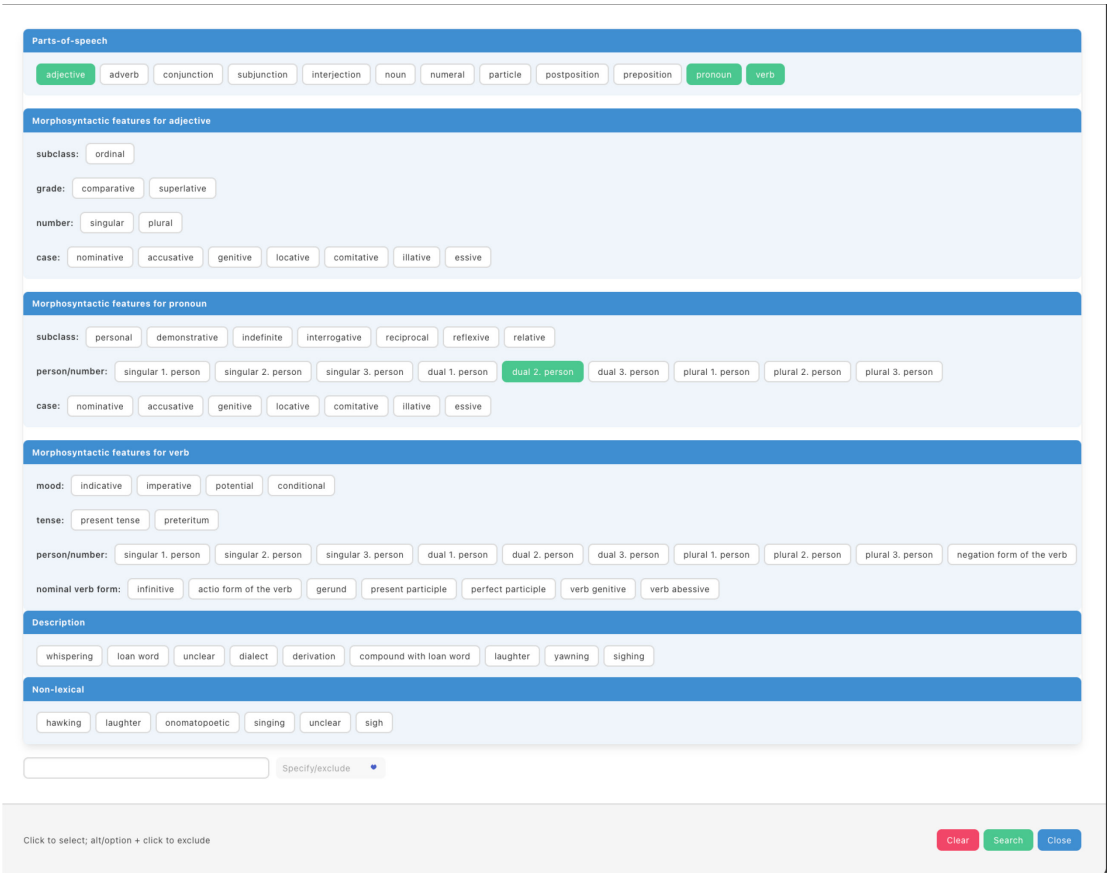
Kart 1: Dei 19 opptaksstadene i LIA Sápmi – Sámegiela hállangiellakorpus

Figur 2 viser Glossa-grensesnittet for enkelt søk i LIA Sápmi. I menyen til venstre kan ein spesifisere søket nærmare på metakategoriar. Dei fleste kategoriane er oppgitt på engelsk, men dei geografiske kategoriane også på nordsamisk, nærmare bestemt «stad» (*Báiki* versus *Place*), «kommune» (*Gielda-Suohkan* versus *District*) og «fylke» (*Fylka* versus *County*).



Figur 2: Grensesnitt for enkelt søk i *LIA Sápmi – Sámegiela hállangiellakorpus*.

Transkripsjonane i LIA Sápmi er morfologisk tagga med den regelbaserte analysatoren Giella-sme (https://giellalt.github.io/lang-sme/, sjå Antonsen & Trosterud 2017), som for LIA-prosjektet vart bygd ut til å kunne analysere dei mange unormerte lånorda som finst i LIA-korpuset (Antonsen 2021). Om ein vel grensesnittet for utvida søk («Extended»), er det mogleg å søke på både ordklassar og morfosyntaktiske kategoriar. Figur 3 viser kva val ein får når ein først har valt ordklassene adjektiv, pronomen og verb. Figuren viser også kva ikkje-lingvistiske taggar ein kan søkje på. Undersøkingane presenterte i Antonsen (2021) og Bentzen (2021) byggjer på søk i LIA Sápmi.



Figur 3: Søkbare morfologiske og ikkje-lingvistiske kategoriar i LIA Sápmi – Sámegiela hállangiellakorpus.

Ytterlegare funksjonar ved Glossa-grensesnittet slik som konkordansar og kopling mellom transkripsjon og lyd vil gå fram av presentasjonen av LIA norsk i følgjande avsnitt.

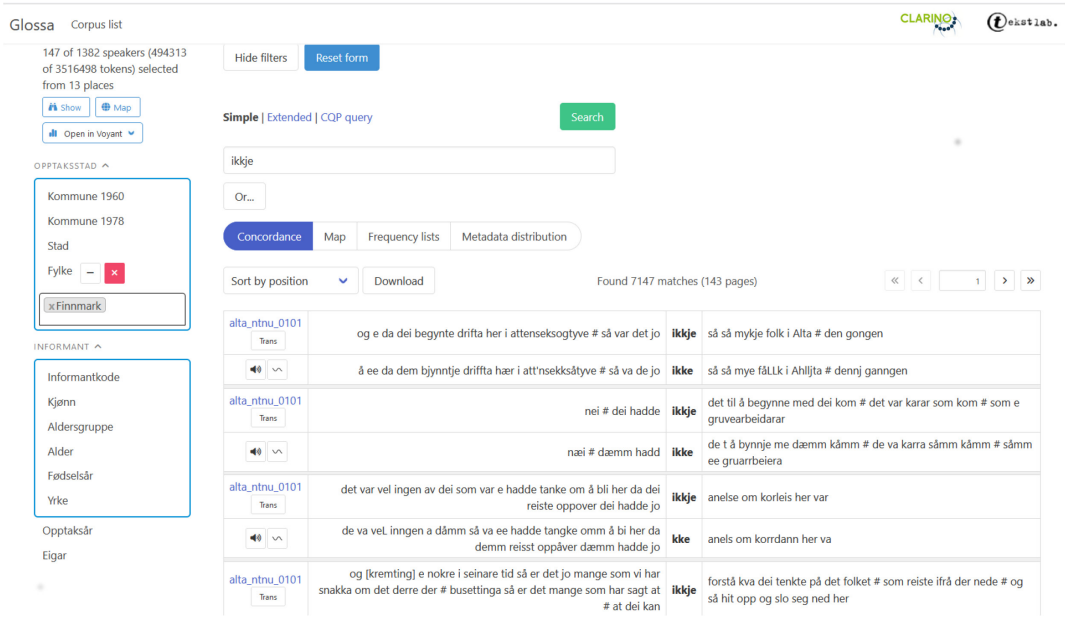
8. LIA norsk - korpus av eldre dialektopptak

LIA norsk er det største korpuset frå LIA-prosjektet. Det inneheld opptak med 1274 informantar frå 226 kommunar, alt i alt om lag 3,5 millionar ord og skiljeteikn. Alle orda er transkriberte på to måtar: ein talemålsnær og ein ortografisk til nynorsk (jf. ovanfor). Vi starta med å transkribere den talemålsnære versjonen, deretter vart den ortografiske laga ved hjelp av ein halvautomatisk translitterator, Oslo-translitteratoren, der den translittererte teksten vart korrekturlesen, og translitteratoren trena fleire gonger for kvar dialekt. Dette sparte prosjektet for mykje tid. Dei to transkripsjonane er knytte saman, og det er mogleg å søke i begge.

Det vart digitalisert over 3000 norske filer i prosjektet. Det utgjer ei større mengde enn det var mogleg å transkribere og leggje inn i korpuset innanfor ramma for LIA-prosjektet. Vi forsøkte å velje ut dei opptaka som hadde best kvalitet samtidig som vi ønskte så stor geografisk spreiing som mogleg. Vi ønskte også fri

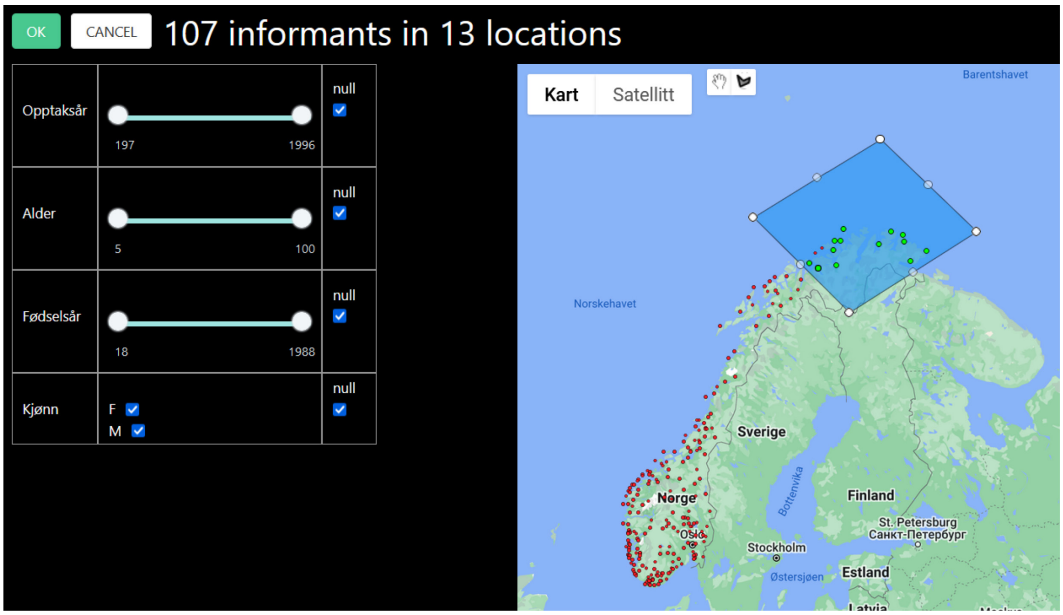
tale og ikkje opplesing av ordlister og liknande som det også fanst mykje av. Det eldste opptaket er frå 1937, det yngste frå 1996. Prosjektet hadde ikkje mandat til å gjere nye opptak eller leite etter opptak frå andre kjelder enn arkiva til dei fire samarbeidsuniversiteta. Det gjorde at vi måtte ta det som fanst, og eit resultat av det er at korpuset har vorte noko ubalansert geografisk sett. Frå Vestfold fylke er det til dømes berre 5318 ord og skiljeteikn, medan det frå Akershus var 31 858 ord og skiljeteikn. Troms og Hordaland har over 500 000 ord og skiljeteikn kvar. I Hagen et al. (2021) er det sju artiklar som i hovudsak baserer seg på undersøkingar av LIA norsk-korpuset og som såleis eksemplifiserer korleis korpuset kan utnyttast.

Figur 4 viser søkjesida for LIA norsk. I søkjeboksen kan ein velje mellom enkelt søk eller meir avanserte søk. Vi kjem tilbake til avansert søk nedanfor. I metadatamenyen til venstre på søkjesida kan ein sjå og velje metadata. Øvst til venstre kan ein få eit oversyn over dei informantane som er valde (*Show*), eller opne metadatautvalet i teksthandsamingsverktøyet *Voyant*. Ved å klikke på *Map*, kan ein velje metadata ved hjelp av kart, sjå figur 5. Resultatet av eit lingvistisk søk blir vist som ein konkordanse. Til venstre for konkordansen kan ein klikke på informantnamnet og få opp all informasjon om informanten. Det er også mogleg å få resultatet omsett ved hjelp av Google Translate. Kvaliteten på og nytten av desse omsetjingane har ikkje vore validert. Eit klikk på lydsymbolet speler av resultatet, og bølgjesymbolet vil gi bølgjeform og spektogram av søkeresultatet.



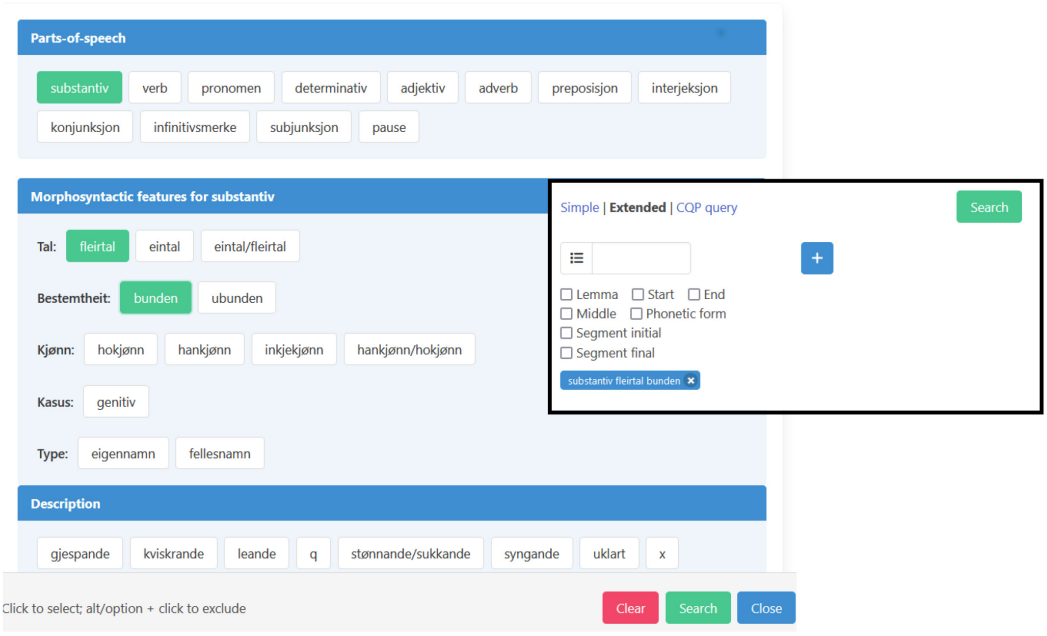
Figur 4: Søkjeside for LIA norsk med konkordans.

Det er også mogleg å velje metadata ut frå eit kart slik det er vist i figur 5. Der kan ein spesifisere metadataa nærmare med skyveknappane i menyane til venstre.



Figur 5: Val av metadata i kart i Glossa 4.0.

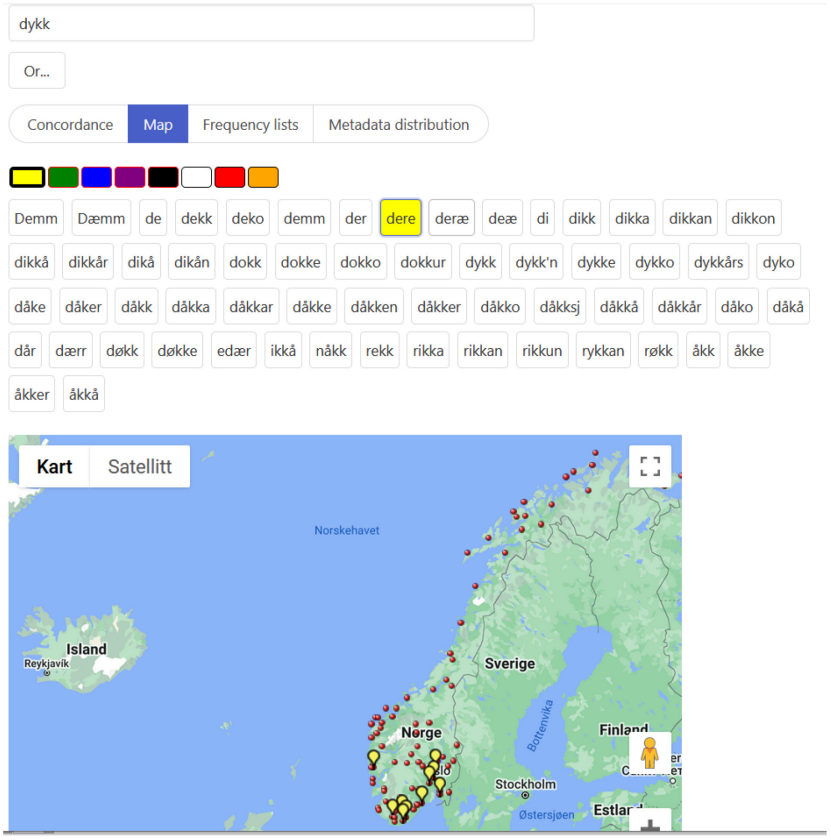
Grensesnittet for eit avansert søk er vist i figur 6. Vindauget til høgre viser søkjeboksen når *Extended* er vald. Her kan ein klikke av for lemma, start og slutt på ord, fonetisk form osv. Klikkar ein på meny–symbolet til venstre for ordsøkjeboksen, får ein opp ordklassemenyen som er vist til venstre i figuren. Eit klikk på ein ordklasse her genererer fleire morfologiske val. I Extended-vindauget til høgre (i blått under avkryssingsboksane) ser ein korleis substantiv, fleirtal, bunden form er valde i ordklassemenyen.



Figur 6: Utvida søk på morfologiske kategoriar i LIA norsk med den nyaste versjonen av Glossa.

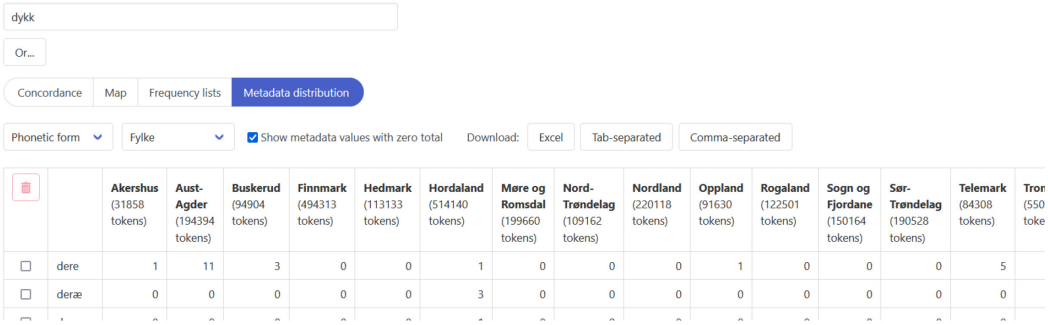
Resultatet av eit søk kan visast på fleire måtar. I tillegg til konkordansevisninga i figur 4, kan ein få ut resultata som kart, frekvensliste og fordelt på metadata. Sidan korpuset har to variantar av kvart ord – ein

talemålsnær og ein ortografisk – kan ein søkje på det ortografiske ordet og sjå distribusjonen av dei ulike dialektformene anten i eit kart eller ein tabell fordelt på metadatakategoriar. Kartvisninga er illustrert i figur 7 for eit søk etter ordforma *dykk* i (standard) nynorsk. Alle uttaleformene er viste som boksar over kartet, og desse kan ein velje å fargeleggje frå ein palett med åtte fargar som i neste omgang syner att på kartet som boblar i den aktuelle fargen. I figuren er boksen for uttaleforma *dere* markert med gult, og i kartet ser ein kva stader korpuset har døme på denne forma frå.



Figur 7: Søk og resultatvisning for den nynorske ordforma dykk.

I figur 8 er det same søket vist med resultatvisning i form av ein tabell fordelt på metadatakategoriar.



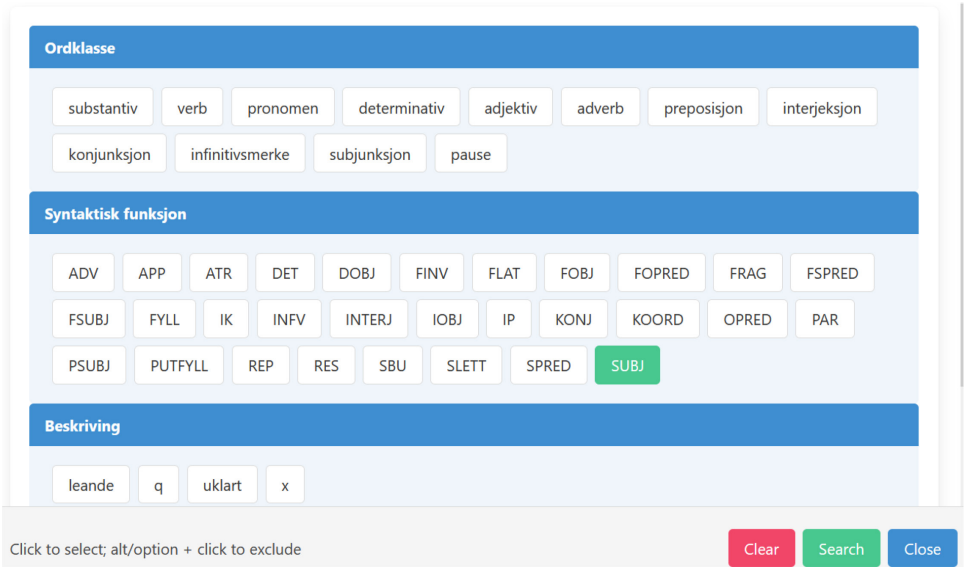
Figur 8: Figuren viser søket på “dykk” distribuert på metadata, her talemålsnær form fordelt på fylke.

9. LIA-trebanken

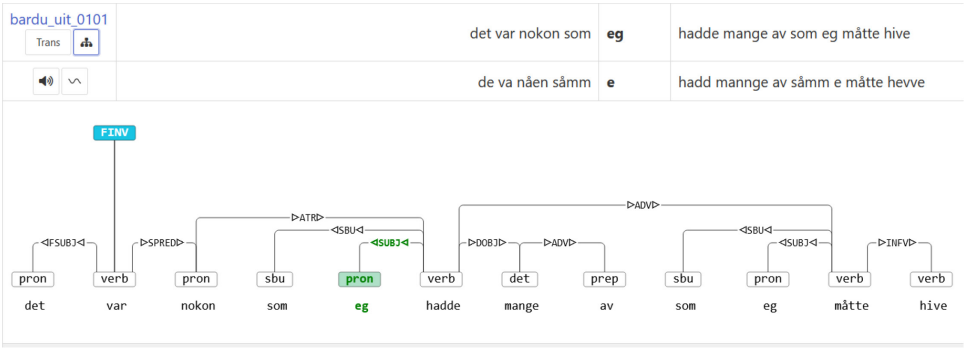
LIA-trebanken inneheld eit utsnitt av lyd og transkripsjonar frå LIA norsk. Det er 55 846 token frå 15 ulike stader som er trekte ut og annoterte med både morfologisk og syntaktisk informasjon. Transkripsjonane er først annoterte med automatiske verktøy utvikla for skriftspråk, deretter er dei manuelt korrigerte av minst ein person. LIA-trebanken er ein dependenstrebank der annotasjonen følgjer retningslinjene i Norsk dependenstrebank for skriftspråk (Kinn et al. 2013). Det er laga eit eige tillegg med retningslinjer for typiske talemålstrekk som pauser, repetisjonar og sjølvkorreksjonar (sjå heimesida til LIA).

LIA-trebanken er søkbar i Glossa, og det er mogleg å søkje på enkle syntaktiske funksjonar, sjå figur 9, i tillegg til søk på ord, ordklasser og morfologiske trekk. Resultata kan dessutan visast som syntaktiske tre som i figur 10.

Trebanken er også nedlastbar i conllx-format, og denne versjonen er brukt til å trene ein parser for talespråk transkribert til nynorsk: LIA-parseren (Øvrelid et al. 2018).



Figur 9: Søkbare syntaktiske variablar i LIA-trebanken – her kryssa av for 'subjekt'



Figur 10: Dependens-tre generert frå eit søk etter syntaktiske variablar i LIA-trebanken.

9. Avslutning

LIA-prosjektet har gjort eldre talemålsopptak for norsk og (nord)samisk tilgjengelege for forsking på ein heilt annan måte enn tidlegare. For det norske korpuset sin del (LIA norsk) representerer dette ein viktig ressurs for ulike former for dialektologiske undersøkingar, og det opnar moglegheiter for å gjere systematiske diakrone undersøkingar med ei viss djupn på ulike språklege variablar for eksempel om ein saman-

liknar med den norske delen av Nordisk dialektkorpus (sjå t.d. Stjernholm & Ims 2021 og Vangsnes & Westergaard 2021).

Det samiske korpuset (LIA Sápmi) har ein langt meir beskjeden storleik enn det norske og omfattar berre nordsamisk, men med tanke på at det førebels ikkje finst noko anna allment tilgjengeleg korpus for samisk talespråk, er det ein nærmast uvurderleg ressurs for undersøkingar av språkleg variasjon innanfor nordsamisk. Når fleire av dei dialektane som er dokumenterte gjennom opptaka, no anten er borte eller svært marginaliserte, gjer det korpuset enno meir verdifullt. Ein kan også håpa på at LIA Sápmi kan bli ein start på noko større og at fleire opptak og fleire samiske varietetar vert lagt til etter kvart, anten i ressursen slik han ligg føre no eller at korpuset kan gå inn i eit nytt og større framtidig samisk talespråkskorpus.

Korpusa som har vore utvikla i LIA-prosjektet, inngår no i porteføljen av digitale ressursar ved einingа Humit som vart oppretta i januar 2023 ved Det humanistiske fakultet, Universitetet i Oslo. Tekstlaboratoriet er ein del av denne nye einingа.

Krediteringar

Det er særs mange som har vore involverte i LIA-prosjektet i større eller mindre omfang, alt frå transkribørar ved dei fire universiteta til teknisk stab og dagleg leiing ved Tekstlaboratoriet, Universitetet i Oslo. Sjå ei oversikt over alle her: http://www.tekstlab.uio.no/LIA/prosjekt.html.

Styringsgruppa for prosjektet vart leidd av Janne B. Johannessen og var elles sett saman av Gjert Kristoffersen frå UiB med Helge Sandøy som vara, Øystein A. Vangsnes frå UiT, Tor A. Åfarli frå NTNU og Svein Arne Solbakk frå Nasjonalbiblioteket med Lisbeth Johannessen som vara.

Ut over ein stor takk til alle som har vore involverte i prosjektet, vil vi også takke to anonyme fagfellar for nyttige tilbakemeldingar på tidlegare versjonar av denne artikkelen.

Referansar

Antonsen, Lene 2021: ‘Lei niogtredve go byggiimet.’ Om unormerte lån fra norsk i samisk talespråk. I Hagen et al., s. 179–200.

Antonsen, Lene og Trond Trosterud 2017: Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk. *Norsk lingvistisk tidsskrift* 35:2, s. 153–185.

Bentzen, Kristine. 2021. VO – OV-variasjon i nordsamisk. Hva kan LIA Sápmi fortelle oss? I Hagen et al., s. 201–216. https://doi.org/10.5617/osla.8483

Hagen, Kristin, Gjert Kristoffersen, Øystein A. Vangsnes og Tor A. Åfarli. 2021. *Språk i arkiva. Ny forsking om eldre talemål frå LIA-prosjektet*. Novus Forlag, ISBN 9788283900811, http://omp.novus.no/index.php/novus/catalog/book/19.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor A. Åfarli, og Øystein A. Vangsnes. 2009. The Nordic Dialect Corpus – an Advanced Research Tool. In Jokinen, Kristiina and Eckhard Bick (eds.): *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. NEALT Proceedings Series Volume 4, https://aclanthology.org/W09-4612/

Johannessen, Janne Bondi; Kristin Hagen; André Lynum og Anders Nøklestad. 2012. OBT+stat. A combined rule-based and statistical tagger. In Andersen, Gisle (ed.): *Exploring Newspaper Language. Corpus compilation and research based on the Norwegian Newspaper Corpus.* John Benjamins Publishing Company, 51–65. https://doi.org/10.1075/scl.49.03joh

Kinn, Kari, Per Erik Solberg og Pål Kristian Eriksen, 2013. *Retningslinjer for morfologisk og syntaktisk annotasjon i Norsk dependenstrebank*. https://tekstlab.uio.no/LIA/pdf/retningslinjer_NDT_norsk.pdf eller på engelsk: https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-10/

Kåsen, Andre, Kristin Hagen, Anders Nøklestad, Joel Priestley, Per Erik Solberg og Dag Trygve Truslew Haug. 2022. The Norwegian Dialect Corpus Treebank. I Nicoletta Calzolari et al.: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, https://aclanthology.org/2022.lrec-1.516/

Nøklestad, Anders og Åshild Søfteland 2007. Tagging a Norwegian Speech Corpus. In *NODALIDA 2007 Conference Proceedings*, https://aclanthology.org/W07-2436/
Stjernholm, Karine og Ingunn Indrebø Ims. 2021. Språkendring i Vika: En komparativ analyse av data fra to talespråkskorpus. I Hagen et al., 2021, s. 109–128.
Vangsnes, Øystein A. og Marit Westergaard. 2021. *Ka LIA fortæll?* Eit gjensyn med kv-spørsmål i norske dialektar I Hagen et al., 2021, s. 155–178.
Øvrelid, Lilja Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg og Janne Bondi Johannessen. 2018. The LIA Treebank of Spoken Norwegian Dialects. I Nicoletta Calzolari et al.: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, https://aclanthology.org/L18-1710/

Nettressursar

CANS - amerikanordisk talespråkskorpus v.3.1: http://www.tekstlab.uio.no/norskiamerika/index.html
CLARINO+: https://clarin.w.uib.no/
Humit: https://www.hf.uio.no/tjenester/humit/
LIA-trebanken: http://www.tekstlab.uio.no/LIA/trebank.html
LIA fildepot: http://www.tekstlab.uio.no/LIA/fildepot.html
LIA norsk: http://www.tekstlab.uio.no/LIA/norsk/index.html
LIA Sápmi - Sámegiela hállangiellakorpus: http://tekstlab.uio.no/LIA/samisk/index.html
Norwegian across the americas: https://www.uib.no/en/lle/134611/norwegian-across-americas
Oslo-translitteratoren: https://www.hf.uio.no/iln/om/organisasjon/tekstlab/tjenester/oslo-translitterator/
ScanDiaSyn-prosjektet og Nordisk dialektkorpus: http://www.tekstlab.uio.no/scandiasyn/
TAUS v.3 - Talemålsundersøkelsen i Oslo: http://www.tekstlab.uio.no/nota/taus/