Novus forlag · eISBN 978-82-8390-082-8

Om LIA-prosjektet og artiklane i denne boka

Kristin Hagen, Øystein A. Vangsnes og Tor A. Åfarli

Denne boka inneheld 13 fagartiklar som med eitt unntak har sine opphav i føredrag frå LIA-prosjektet sitt sluttseminar i Trondheim 26.–27. november 2019. Arbeida tar for seg vidt forskjellige emne på grunnlag av gamle talemålsopptak som no er gjort tilgjengelege gjennom LIA-prosjektet og tre nye og brukarvenlege elektroniske korpus: 1) LIA norsk – korpus av eldre dialektopptak, 2) LIA Sápmi – Sámegiela hállangiellakorpus, 3) CANS – amerikanordisk talespråkskorpus v.3.1.

Om LIA-prosjektet

LIA står for Language Infrastructure made Accessible, eit stort infrastrukturprosjekt finansiert gjennom forskingsinfrastrukturprogrammet til Noregs forskingsråd (NFR) i perioden 2014–2019. Det var eit nasjonalt samarbeidsprosjekt mellom Universitetet i Oslo, Universitetet i Bergen, UiT Noregs arktiske universitet, Noregs teknisk-naturvitskaplege universitet, Norsk Ordbok 2014 og Nasjonalbiblioteket. Initiativtakar til prosjektet var professor Janne Bondi Johannessen, og det har heile tida vore forankra ved Tekstlaboratoriet, UiO. Det vart leidd av Janne saman med ei styringsgruppe med representantar for dei ulike institusjonane i prosjektet.

Det overordna formålet med LIA-prosjektet var å redde eldre talemålsopptak med norsk og samisk språk som levde eit bortgøymt

og i stor grad bortgløymt tilvære i arkiv og instituttbibliotek ved dei «gamle» universiteta våre i Oslo, Bergen, Trondheim og Tromsø, og somme også på private hender. Vidare var det eit mål å transkribere og annotere ei stor mengd av dei viktigaste av desse opptaka og leggje dei inn i databasar slik at dei kunne revitaliserast som verdifullt forskingsmateriale. Opptaka vart digitaliserte ved Nasjonalbiblioteket i Mo i Rana, og kopiar er langtidslagra der. Prosjekttilsette ved dei fire universiteta har gått gjennom opptaka, katalogisert dei, utstyrt dei med metadata, og ein stor del av materialet har også vorte transkribert.

Dei transkriberte opptaka har vorte samla i dei tre korpusa nemnde ovanfor, som alle er elektronisk tilgjengelege og søkbare med det brukarvenlege grensesnittet Glossa, utvikla ved Tekstlaboratoriet. Meir om korpusa følgjer nedanfor. Gjennom prosjektet vart også den såkalla B-serien frå Talemålsundersøkelsen i Oslo (TAUS) transkribert og lagt inn i ein ny versjon av det eksisterande TAUS-korpuset (v.3). I tillegg har det vorte oppretta eit søkbart fildepot som inneheld alt digitalisert materiale på norsk, også det som ikkje er transkribert. Her er det mogleg å lytte til lydfilene og å be om å få laste ned filer som ikkje har sensitivt innhald.

For språkteknologiske og andre formål er det laga ein nedlastbar versjon av LIA norsk som omfattar 553 transkripsjonar i tekstformat med tilhøyrande lydfiler. Det er også mogleg å laste ned LIA-trebanken, ein trebank som inneheld 5 250 talemålssegment med 55 410 ord og skiljeteikn frå LIA norsk, og som er annotert med morfologisk og syntaktisk informasjon. Annoteringa er gjort maskinelt, men alt er gjennomgått og korrigert manuelt.

Med dette arbeidet har LIA-prosjektet tilgjengeleggjort datamateriale frå fleire tiår med dialektologisk innsamlingsarbeid utført av fleire generasjonar målføregranskarar ved universiteta våre og av andre. Som nemnt ovanfor var det overordna formålet med prosjektet å redde dei eldre talemålsopptaka. Dei fleste av desse opptaka var nemleg lagra på magnetband, og det var grunn til å frykte at dei etter kvart ville gå tapt. LIA-prosjektet var derfor ikkje berre eit lingvistisk prosjekt, men også eit viktig kulturhistorisk redningsprosjekt. Det er fortenesta til Janne at ho såg begge desse aspekta og hadde hand-

lekraft til både å initiere LIA-prosjektet og leie det gjennom heile prosjektperioden og fram til ei vellukka avslutning.

LIA-prosjektet har også vore eit vellukka samarbeidsprosjekt med mange spennande diskusjonar og prosjektmøte ved dei fire universiteta, i Mo i Rana og til sist ein prosjekttur til Finnmark avslutta med seminar på Sámi allaskuvla (Samisk høgskole) i Kautokeino i september 2018. Prosjektet har dessutan hatt til saman fleire titals prosjekttilsette ved dei fire samarbeidsuniversiteta, der mange av dei har vore deltidstilsette studentar som har transkribert, korrekturlese og annotert. Vi er svært takksame for kvar og ein som har bidratt til prosjektet.

Om arbeidet med denne boka

Denne boka er den tredje artikkelsamlinga på Novus forlag med arbeid basert på norske og nordiske talespråkskorpus utvikla ved Tekstlaboratoriet. I praksis, om ikkje formelt, dannar dei tre bøkene ein serie. Den første boka med tittelen *Språk i Oslo: Ny forskning omkring talespråk* redigert av Janne Bondi Johannessen og Kristin Hagen kom i 2008 og inneheld 19 artiklar frå eit seminar i 2006 knytt til det då nyleg ferdigstilte NoTa-korpuset med talemålsopptak frå Oslo samla inn på byrjinga av 2000-talet. I 2014 kom *Språk i Norge og nabolanda: Ny forskning om talespråk*, også redigert av Janne Bondi Johannessen og Kristin Hagen, med 13 artiklar frå eit seminar året før knytt til ferdigstillinga av Nordisk dialektkorpus, eit av resultata frå det store nordiske samarbeidsprosjektet *Nordisk dialektsyntaks*.

Det var meininga at også denne boka, *Språk i arkiva: Ny forsking om eldre talemål* frå LIA-prosjektet med bidrag frå avslutningsseminaret for LIA-prosjektet i november 2019, skulle redigerast av Janne Bondi Johannessen og Kristin Hagen. Men våren 2020 vart Janne akutt verre av kreftsjukdomen ho hadde kjempa mot i lengre tid, og 15. juni 2020 døydde ho. I hennar stad gjekk då Tor A. Åfarli, Gjert Kristoffersen og Øystein A. Vangsnes inn som redaktørar. Alle tre

hadde vore med i styringsgruppa for prosjektet som representantar for høvesvis NTNU, UiB og UiT.

Våren 2021 ramma eit nytt dødsfall arbeidet med boka då Gjert Kristoffersen mista livet i ei trafikkulukke 29. mai. Gjert hadde fram til då lagt ned ein stor og uvurderleg innsats med boka. I tillegg til å ha hovudansvar for nokre av artiklane, hadde han også formatert alle bidraga i tråd med den malen som skulle nyttast.

I tillegg til at dei tre bøkene alle har sprunge ut av forskings- og infrastrukturprosjekt med Janne Bondi Johannessen som primus motor, har alle omslaga på bøkene vorte teikna av sonen hennar, Edvard Bondi Knowles. I 2008 var han ein teikneglad tenåring, i 2014 student i landskapsarkitektur ved NMBU og no i 2021 utøvande landskapsarkitekt og dessutan aktiv jazzmusikar. Vi er svært glade også for at dette knyter dei tre bøkene saman.

Om dei tre LIA-korpusa som er nytta i artiklane i boka

Det er empirisk materiale frå dei tre LIA-korpusa nedanfor som er nytta i dei 13 fagartiklane som du finn i denne boka. Samla sett demonstrerer dei kor rikt materialet frå LIA-prosjektet er som utgangspunkt og empirisk grunnlag for teoretisk analyse og språkforskinga generelt. Det er likevel viktig å vere klar over at LIA-korpusa ikkje er balanserte korpus med omsyn til utval av informantar. Sidan LIA-prosjektet har arbeidd med eldre talemålsopptak, har vi måtta bruke det som fanst av opptak, og det har ført til ubalanse, både med omsyn til variablar som kjønn, bustad og alder.

Prosjektet har lagt vinn på å finne opptak av best mogleg lydkvalitet frå flest mogleg stader. Vi har også prioritert opptak med fri tale og ikkje opplesing av tekst eller oppramsingar av bøyingsparadigme eller stadnamn. Slik har korpusa også vorte gullgruver av kulturhistorisk verdi med tema som matlaging, tømmerdrift, draktskikkar osv.

LIA norsk – korpus av eldre dialektopptak

LIA norsk inneheld opptak og transkripsjonar frå fire universitet: NTNU, UiB, UiO og UiT. Korpuset inneheld også eit delmateriale frå Målførearkivet ved UiO som tidlegare var å finne i Nordisk dialektkorpus.

LIA norsk har 1374 informantar frå 222 kommunar og inneheld ca. 3,5 millionar ord. Korpuset er transkribert både talemålsnært og ortografisk til nynorsk og er hausten 2021 morfologisk tagga med ein nyutvikla talemålstaggar for nynorsk. Korpuset er søkbart både med omsyn til ortografisk form, talemålsnær form, lemma og morfologiske opplysningar. Resultata kan filtrerast gjennom metadata som stad, kommune, fylke, kjønn og opptaksår. Det eldste opptaket er frå 1937.

LIA Sápmi – Sámegiela hállangiellakorpus

LIA Sápmi er det første talespråkskorpuset med samiske dialektar. Korpuset har nesten 190 000 ord fordelt på 122 informantar frå 19 stader. Opptaka er frå tidsrommet 1960–1987, og det er opptak frå store delar av det nordsamiske området. Mange opptak stammar frå samlinga til Nils Jernsletten, som var professor i samisk ved UiT.

Opptaka er ortografisk transkriberte, og korpuset har fått automatisk lingvistisk analyse av Giellatekno ved UiT. Det er dermed mogleg å søke på ordklasse og lemma, i tillegg til ordform. Søka kan sjølvsagt filtrerast gjennom metadata.

CANS – amerikanordisk talespråkskorpus v. 3.1

CANS - amerikanordisk talespråkskorpus v. 3.1 er det einaste LIA-korpuset som også har nyare opptak. Informantane er eldre norsk-amerikanarar som lærte norsk heime, dei fleste på 1920- og 30-talet. Opptaka blei gjort i perioden 2010 – 2016 av Janne Bondi Johannessen og andre. Korpuset har også ein liten del amerikasvensk med

opptak frå Ida Larsson og andre (2011 – 2014). CANS inneheld også gamle opptak frå Didrik Arup Seip og Ernst W. Selmer (1931), Einar Haugen (1942) og Arnstein Hjelde (1987, 1990, 1992). Arbeidet med korpuset starta før LIA-prosjektet fekk klarsignal, men blei mange gonger større på grunn av LIA-midlane.

Korpuset inneheld 268 informantar frå USA og Canada, 22 som snakkar svensk og 246 som snakkar norsk, i alt nesten 775 000 ord. Det er både intervju og samtalar, og opptaka er transkriberte både talemålsnært og ortografisk til bokmål. Transkripsjonane er tagga automatisk med NoTa-taggaren, og korpuset er dermed søkbart på same måten som LIA norsk.

Kort om artiklane i boka

Av dei 13 artiklane i boka tek sju utgangspunkt i LIA norsk, to i LIA Sápmi og fire i CANS, og dei dannar såleis tre tematiske hovudbolkar. Nedanfor gir vi eit kort oversyn over artiklane, inndelte alfabetisk etter namnet til førsteforfattaren innanfor dei tre hovudbolkane. Vi byrjar med dei sju artiklane som er grupperte under temaet norsk talemål.

Den første har tittelen «Palatalisering av velarer: historie, variasjon og normering» og er skriven av **Ivar Berg og Leiv Inge Aa**. Forfattarane undersøkjer variasjonen mellom velar plosiv /g, k/ og palatal frikativ /j, ç/ i innlyd i norske dialektar (jf. ord som *skog(j)en*, *tenk(j)e*). Dei peikar på at variasjonen lett blir utjamna der palatalane inngår i morfofonologisk veksling med ein velar, og at desse palatalane er på retur i dialektane. Dei undersøkjer om LIA-korpuset og Nordisk dialektkorpus kan kaste lys over denne problematikken, og dei ser også på om svekkinga av palatalane i talemålet kan ha samanheng med j-bortfall i nynorsk skrift.

I artikkelen «Ka farsken? Realisering av /r/ som sibilant foran /k/ i nordnorsk» tar **Gjert Kristoffersen og Randi Neteland** for seg bruken av sibilanten [ʂ] som allofon av /r/ i rk-sekvensar i nordnorske dialektar (jf. uttalen av ordet *farsken* i tittelen) med utgangspunkt i LIA-korpuset. Dei finn at sibilantuttale er utbreidd i heile Nord-

Noreg, men at det er store individuelle ulikskapar når det gjeld bruksfrekvens, også blant dei som har mange belegg på variabelen.

Signe Laake og Lilja Øvrelid presenterer i artikkelen «Forskjeller mellom talemål og skriftspråk: Hva kan trebanker fortelle oss?» ein samanliknande studie av norsk talemål og skriftspråk. Dei tar for seg fleire syntaktiske ulikskapar mellom skrift og tale som har blitt undersøkte i tidlegare studiar, som t.d. sideordning vs. underordning og utelating av syntaktiske ledd. I sin studie brukar dei den nyleg utvikla LIA-trebanken og den skriftspråklege Norsk Dependenstrebank og viser korleis desse gjer det mogleg å utføre presise syntaktiske søk som legg til rette for at dei ulike syntaktiske fenomena kan undersøkjast på ein effektiv måte både i tale og skrift.

I artikkelen «LIA-korpuset som ressurs i revisjonen av tre ordbøker» ser **Bente Selback og Terje Svardal** på kva ressurs eit talemålskorpus som LIA norsk kan vere i arbeidet med revisjonen av tre ordbøker: standardordbøkene Bokmålsordboka og Nynorskordboka og dokumentasjonsordboka Norsk Ordbok. I ein ordboksrevisjon arbeider ein både med å finne kva nye ord (lemma) som bør takast inn i ordbøkene, og ein vurderer dei eksisterande ordboksartiklane: Er eit ord så sentralt at det bør vere med vidare? Er definisjonane i tråd med gjengs språkbruk? Kva bruksdøme fungerer best? I eit slikt revisjonsarbeid er det viktig å ha gode kjelder, og i artikkelen vurderer forfattarane kva ein kan og eventuelt ikkje kan bruke LIA-korpuset til i dette arbeidet.

I artikkelen «Språkendring i Vika. En komparativ analyse av data fra to talespråkskorpus» presenterer **Karine Stjernholm og Ingunn Indrebø Ims** ein pilotstudie der dei brukar materiale frå LIA norsk og Nordisk dialektkorpus til å teste ein hypotese om ulik språkutvikling i det vikværske dialektområdet i perioden 1960–2009. Nærmare bestemt undersøkjer dei overgang frå *ær*-suffiks til *er*-suffiks i kategoriane hankjønn, fleirtal, ubestemt, samt nomen agentis og presens av svake verb. Dei finn indikasjonar på at det er ulik bruk av desse suffiksa på tvers av dialektområdet, og dei diskuterer korleis sosiolingvistisk teori kan bida med nyttige perspektiv for å forstå språkutviklinga i området.

I artikkelen «Trykkplasseringa i latinske lånord og partikkelverb i tre norske dialektområde» ser **Eirik Tengesdal og Björn Lundquist** på sambandet mellom trykkplassering i latinske lånord (som *butikk*) og partikkelverb (som *gå ut*), der trykkplasseringa på fyrstestavinga eller sistestavinga varierer. Dei reknar med at partikkelverb viser ein del variasjon som skriv seg frå språkinterne strukturelle faktorar, og at desse faktorane ikkje påverkar latinske ord i like stor grad, og vidare reknar dei med at ei gransking av ikkje-syntaktisk variasjon i latinske ord kan bidra til ei betre forståing av trykkvariasjonen i partikkelverb. Med det som utgangspunkt undersøkjer dei talarar frå Trøndelag, Hedmark og Finnmark og gir ei grundig skildring og analyse av den relevante trykkvariasjonen. Det empiriske materialet hentar dei frå LIA-korpuset og Nordisk dialektkorpus, og dei finn at LIA og NDK er veleigna for prosodisk gransking.

Til slutt i bolken om norsk talemål presenterer **Øystein A. Vangsnes og Marit Westergaard** i artikkelen «Ka LIA fortæll? Eit gjensyn med kv-spørsmål i norske dialektar» ei undersøking som replikerer ei tidlegare korpusbasert undersøking av kv-spørsmål utan V2 i norske dialektar. Den første undersøkinga tok for seg temaet med utgangspunkt i Nordisk dialektkorpus, som har opptak frå perioden 2005–2010 frå 111 stader i landet. Den siste undersøkinga, som er gjennomført for denne artikkelen, tar for seg temaet med utgangspunkt i det norske LIA-korpuset, som inneheld eldre opptak frå 222 ulike stader, men der opptaka er spreidde over ein lengre tidsperiode både med tanke på opptaksår og fødeåret til informantane. Forfattarane finn at data frå LIA-korpuset langt på veg stadfestar det generelle inntrykket frå NDK når det gjeld kv-spørsmål utan V2 i norske dialektar, men ein mogleg skilnad mellom dei to korpusa når det gjeld det vestnorske tilfanget, blir drøfta.

Bolk nummer to er dei to artiklane som tar for seg samisk. Den første, av **Lene Antonsen**, har tittelen «'Lei niogtredve go byggiimet.' Om unormerte lån fra norsk i samisk talespråk». Denne artikkelen undersøkjer materiale annotert som framandspråkleg og sitatlån i nordsamisk i talespråkskorpuset LIA Sápmi, og vurderer dei mest frekvente orda i høve til ordbøker og tekstkorpus. Analysen viser at

ein del av dette materialet er etablerte lånord i munnleg språk, og mange av dei finst også i ordbøker, men orda er fråverande i skriftleg språk. Vidare viser analysen at størstedelen av materialet er spontanlån der det innlånte leksikalske elementet blir tilpassa samisk morfologi og syntaks.

Det andre bidraget i samisk-bolken har tittelen «VO – OV-variasjon i nordsamisk: Hva kan LIA Sápmi fortelle oss?» og er forfatta av **Kristine Bentzen**. Artikkelen vart opphavleg publisert i *Bauta: Janne Bondi Johannessen in memoriam* (*Oslo Studies in Language* 11: 2, 2020), men er tatt med også i denne boka sidan undersøkinga er basert på materiale frå LIA Sápmi. I undersøkinga ser Bentzen på vekslinga mellom VO- og OV-ordstilling i nordsamisk talemål, og ho finn at VO er det mest frekvente mønsteret overordna sett. Men i materialet finst det også mange tilfelle av OV-leddstilling i samband med bruk av hjelpeverb, og i tillegg førekjem det nokre tilfelle der objektet står framfor både hjelpeverb og hovudverb (utan at objektet er tematisert). Forfattaren føreslår ein detaljert formell syntaktisk analyse av dei ulike leddstillingsmønstera.

Tredje og siste bolken omfattar fire bidrag som på ulike vis tar for seg problemstillingar som kan grupperast under temaet amerikanorsk. Den første er forfatta av **Ragnhild Eik og Anu Laanemets** og har tittelen «Å være eller å bli – det er spørsmålet. Ei sammenligning av verbene VÆRE og BLI i amerikanorsk og norgenorsk». Artikkelen baserer seg på data frå korpusa CANS (Corpus of American Nordic Speech) og LIA norsk, og undersøkjer verba VÆRE og BLI brukte saman med predikata *født*, *konfirmert* og *gift* i norsk talt i Amerika samanlikna med norsk talt i Noreg. Undersøkinga tar utgangspunkt i tidlegare studiar av amerikadansk som viser at VÆRE i nokon grad har tatt over for BLI saman med desse predikata, truleg på grunn av kontakten med engelsk. Forfattarane finn noko overraskande at eit tilsvarande resultat ikkje gjeld for amerikanorsk, og dei diskuterer korleis dette resultatet skal forståast.

Neste artikkel i denne bolken er «Argumentplacering i norskt arvspråk i Amerika» av **Ida Larsson og Kari Kinn**. Denne artikkelen undersøkjer distribusjonen til subjekt, objekt og partiklar i eldre og yngre amerikanorsk, basert på CANS-korpuset, og i talespråket i

Noreg omkring tida for masseutvandringa til Amerika, basert på LIA-korpuset. Eit hovudfunn er at leddstillinga varierer i begge datamateriala, og at det delvis er meir variasjon i distribusjonen i amerikanorsk enn i moderne norsk talemål. Vidare finn forfattarane blant anna at variasjonen held seg relativt stabil over fleire generasjonar i amerikanorsk, men likevel slik at subjektskifte og objektskifte blir mindre frekvent over tid, noko dei føreslår å forklare ved å vise til økonomiprinsipp som favoriserer 'uskifta' posisjonar, men dei drøftar også andre faktorar som kan spele inn.

I artikkelen «Adjektivkongruens i amerikanorsk» ser **Brita Ramsevik Riksem, Terje Lohndal og Tor A. Åfarli** på attributive engelske adjektiv i amerikanorske nominalfrasar som har norsk struktur og norsk substantiv, og dei undersøkjer om slike engelske adjektiv syner norsk bøying, basert på data frå det nyare CANS-korpuset. Resultata viser at i dei få tilfella i dette datamaterialet der eit engelsk attributiv adjektiv står saman med eit norsk substantiv, er det ikkje norsk bøying, noko som er uventa, sidan engelske substantiv og verb som blir brukte i ein norsk kontekst som hovudregel får norsk bøying. Artikkelen diskuterer kvifor adjektiva oppfører seg annleis enn substantiv og verb og føreslår ein formell syntaktisk analyse av dette.

Den siste artikkelen i den amerikanorske bolken er «'Ein må tenke seg om, ser du veit du, kva ord ein skal bruke.' Pragmatiske partiklar i amerikanorsk» av **Åshild Søfteland og Arnstein Hjelde**. Denne artikkelen tar for seg bruken av dei pragmatiske partiklane *ser du*, *veit du* og *trur eg* i CANS. Desse blir samanlikna med bruk av engelsk *you see*, *you know*, *I think* og norsk *du veit*, *du ser*, *eg trur* i det same datamaterialet. Artikkelen greier ut om frekvens, syntaktisk posisjon og overordna pragmatisk analyse, og forfattarane drøftar også metodologiske problemstillingar. Hovudfunnet er at talarane av amerikanorsk brukar desse pragmatiske partiklane som støtte for informasjonsflyten i samtalesituasjonen, og at dei kan utnytte kombinasjonar av norske og engelske bruksmønster som ein ressurs i språkbruken.

LIA-lenker:

CANS - amerikanordisk talespråkskorpus v.3.1:
https://tekstlab.uio.no/glossa2/cans3

Fildepot for LIA:
https://lia.tekstlab.sigma2.no/

Glossa:
https://www.hf.uio.no/iln/om/organisasjon/tekstlab/tjenester/glossa/

LIA norsk - korpus av eldre dialektopptak:
https://tekstlab.uio.no/glossa2/lia_norsk

LIA-prosjektet:
https://tekstlab.uio.no/LIA/

LIA-trebanken:
https://tekstlab.uio.no/LIA/trebank.html

LIA Sápmi - Sámegiela hállangiellakorpus:
https://tekstlab.uio.no/glossa2/saami

TAUS v.3:
https://tekstlab.uio.no/glossa2/taus3