Norsk Lingvistisk Tidsskrift · Årgang 35 · 2017 153–185

Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk

Lene Antonsen og *Trond Trosterud*

Artikkelen presenterer et analyseprogram for nordsamisk løpende tekst, bestående av en morfologisk transduser og en føringsgrammatikk for disambiguering og syntaktisk analyse. Artikkelen drøfter hva programmet kan fortelle oss om nordsamisk grammatikk når det brukes til å analysere et nordsamisk tekstkorpus bestående av til sammen 25 millioner ord. Vi ser på produktiviteten for en del sentrale orddanningsprossesser for substantiv og verb. Sammensetning er langt mer produktiv enn ordavledning, og av de vanligste sammensetningstypene er forledd i nominativ entall den klart mest produktive. Vi bruker også analyseprogrammet for å måle ulike lingvistiske parametre for substantiv og finitte verb, og undersøker hvor sjangeravhengig disse trekkene er. Trass i størrelsen er korpuset ikke godt balansert for å forske på finitte bøyingskategorier, og slik forsking må ta hensyn til dette. Derimot er det liten variasjon mellom sjangerne når det gjelder substantivenes kasusfordeling.

Nøkkelord: morfologi; nordsamisk; korpuslingvistikk; sammensetning; produktivitet; NLP

1 Introduksjon[1]

Artikkelen presenterer *Giella-sme*,[2] et analyseprogram for nordsamisk, og bruker det til å analysere ulike aspekt ved samisk ordstruktur. Ana-

1. Takk til vår kollega Ciprian Gerstenberger for uvurderlig hjelp med tilrettelegging av korpusdata, og våre kolleger i Divvun-gruppa ved UiT Norges arktiske universitet for arbeid med Giella-sme, og med innsamling og konvertering av korpus.

2. *giella* er nordsamisk for 'språk', og navnet på infrastruktur og analyseprogram for bortimot 50 språk ved UiT, jf. http://giellatekno.uit.no/doc/lang/index.html. *sme* er iso-kode 639-2 for nordsamisk.

lyseprogrammet utgjør også en modell av nordsamisk morfofonologi (eller snarere morfografemikk, i og med at modellen analyserer skriftspråk), morfologi og syntaks. Morfologi og morfofonologi danner et integrert hele, mens den syntaktiske delen er bygd opp etter svært ulike prinsipper.

Den morfologiske delen av analyseprogrammet bygger på eksisterende ordbøker og analyse av tekst, og inneholder dermed leksikaliserte[3] ordavledninger og sammensetninger. Samtidig er det i stand til å analysere de samme orddannningsprosessene dynamisk, og gir dermed ofte to analyser av samme ord, en analyse som viser den leksikaliserte forma, og en som viser delene den leksikaliserte forma består av. Slik vil man kunne se ordene både innafra (derivasjonelt) og utafra (morfosyntaktisk). Analyseprogrammet er en del av en tradisjon innafor automatisk analyse av naturlige språk som kombinerer dyp grammatisk basert analyse med robust parsing (høy presisjon for analyse av store tekstsamlinger), til forskjell fra alternative modeller, som legger vekt på den ene eller andre av disse to prioriteringene.

Artikkelen drøfter hva *Giella-sme* kan fortelle oss om nordsamisk grammatikk når det brukes til å analysere SIKOR (Sámi Internationála KORpus), et nordsamisk tekstkorpus bestående av til sammen 25 millioner ord. Vi bruker analyseprogrammet for å måle ulike lingvistiske parametre for substantiv og finitte verb. Halvparten av tekstene i SIKOR er avistekster, og de skjønnlitterære tekstene utgjør bare 2 % av korpuset. Vi ser på hvilke av disse lingvistiske trekkene som er sjangeravhengig og hvilke som ikke er det.

I del 2 presenterer vi det lingvistiske og språkteknologiske rammeverket vi arbeider innafor, og i del 3 blir SIKOR presentert. I del 4 presenterer vi den morfo(fono)logiske modellen av nordsamisk og viser hvordan vi kan bruke den til å se på ordene innafra, dvs. se på sammensetninger og ordavledninger i korpuset. Vi ser på hvilken sammensetningstype som er mest produktiv, og sammenligner produktiviteten med noen sentrale avledningsprosesser. I del 5 beskriver vi den syntaktiske delen av analyseprogrammet, som velger den riktige morfologiske analysen i kontekst. Vi ser på i hvor stor grad noen egenskaper ved verb og substantiv framstår forskjellig i de forskjellige sjangrene i korpuset. I siste del kommer en oppsummering og konklusjon.

3. Med "leksikalisert" mener vi i denne artikkelen "lagt til i lista over leksem i transduseren". Jf. avsnitt 4.2 for ei drøfting av leksikalisering i den nordsamiske transduseren.

2 Bakgrunn

Den datalingvistiske modellen for nordsamisk grammatikk er laget i tradisjonen etter Koskenniemi (1983, jf. Karttunen og Beesley (2005) for et historisk overblikk). Hans grammatiske modell for finsk ble lagd som to automater, den ene for finsk konkatenativ morfologi (suffiksering), og den andre for de morfologiske og morfofonologiske prosessene ordformene går gjennom i løpet av bøyingsprosessen. For finsk inkluderer dette stadieveksling, diftongforenkling, stammeveksling og vokalharmoni, nordsamisk har de samme prosessene (bortsett fra vokalharmoni), og i tillegg en rekke endringsprosesser for stammevokal og -konsonant.

Den første generasjonen av modeller av denne typen ble laget av kommersielle firma, og tatt i bruk i språkteknologiske applikasjoner, men ikke gjort allment tilgjengelig. Det inkluderer modeller for finsk og for de nordiske språka (ved firmaet Lingsoft, jf. Arppe 2005). Tilsvarende modeller ble laget for de fleste større europeiske språk, og for tyrkisk, koreansk og japansk, av Xerox (Karttunen 2000). Etter 2010 har det ved UiT Norges arktiske universitet (UiT) blitt utviklet en språkuavhengig infrastruktur, som i utgangspunktet inneholdt fullskalamodeller for 10 ulike sirkumpolare språk, (jf. Moshagen m.fl. 2014).

Til den morfologiske modellen blir det for de fleste formål brukt en syntaktisk komponent, en føringsgrammatisk modell (eng. *constraint grammar*) i tradisjonen etter Fred Karlsson (Karlsson 1990, Karlsson m.fl. 1995). Dette er også tilfelle for nordsamisk, og den nordsamiske føringsgrammatikken blir presentert i del 5. Mens nordsamisk morfologi er modellert som en endelig tilstandstransduser, eller som en regulær grammatikk i Chomskyansk forstand, bygger ikke den syntaktiske delen av Giella-sme på en tilsvarende syntaktisk modell. I den grad generativ grammatikk har blitt forsøkt brukt i automatisk analyse, har det vært i form av kontekst-frie grammatikker. Disse fungerer som filter som slipper gjennom de og bare de setningene som kan genereres av regelsettet satt opp i disse grammatikkene. Selv om mye arbeid har blitt lagt ned i å skrive slike grammatikker, har de aldri resultert i robuste modeller for grammatisk analyse av løpende tekst, og den syntaktiske modellen for samisk bruker i stedet føringsgrammatikk (eng. 'constraint grammar').

Føringsgrammatikken bruker et sett av *føringer* (eng. 'constraints') for i hvilken kontekst hver type analyse kan opptre. Både regelformalisme og kompilatorer har blitt videutviklet av Tapanainen (1996) og seinere av Eckhard Bick og andre (jf. Visl-Group 2008), som er den som brukes for samisk. Det er laget føringsgrammatikker for flere titalls språk, for eksempel for finsk (Karlsson 1990), engelsk (Karlsson m.fl. 1995), norsk (Johannessen m. fl. 2012) og

portugisisk (Bick 2000). Disse grammatikkene er i bruk i ulike praktiske applikasjoner, for eksempel i grammatikkontrollprogram for Microsoft Office[4], i maskinoversetting[5], og i analyseprogram som blir brukt for å produsere gullkorpora for statistiske modeller[6].

3 Det nordsamiske korpuset

Det samiske korpuset SIKOR er et elektronisk tilgjengelig korpus for seks samiske og flere andre uralske språk. Den nordsamiske delen inneholder 25 millioner ord, samlet inn ved UiT Norges Arktiske universitet. Korpuset er åpent tilgjengelig for korpussøk på internett[7], og i underkant av halvparten er tilgengelig for nedlastning under en fri lisens. For en oversikt over metodologien ved innsamlinga, se Huhmarniemi m.fl. (2007).

SIKOR er satt sammen av en stor del av all elektronisk tilgjengelig nordsamisk tekst (innsamlet dels direkte fra institusjoner som har produsert samisk tekst, og dels fra internett). Nordsamisk blir ofte betraktet som et minoritetsspråk med svært få ressurser, og sammenlignet med f.eks. svensk, med 9,23 milliarder ord tilgjengelig i den svenske språkbanken[8], er det selvfølgelig et lite korpus. Likevel er SIKOR like stort som korpussamlingene for større språk var for et par tiår siden, og langt større enn det tidlige balanserte korpora var. Det banebrytende Brown-korpuset (Kučera og Francis 1967, et korpus utgitt ved Brown University som bestod av et representativt utvalg av all tekst utgitt på engelsk i USA i 1961), inneholder f.eks. bare en million ord. Siste publiserte versjon av det mest kjente korpuset over et urfolksspråk, Nunavut Hansard, et korpus bestående av en oversettelse til inuktitut av diskusjonene i den kanadiske nasjonalforsamlinga, inneholder 2,6 millioner ord (for perioden 1999–2008, en oppdatert versjon vil anslagsvis være dobbelt så stor)[9].

4. https://www.lingsoft.fi/tuotteet/office for svensk og finsk, norsk bokmål og dansk språk er integrert i Microsoft sine produkter.
5. http://gramtrans.com inneholder f.eks. maskinoversettelsesprogrammer for tekstproduksjon, fra engelsk, portugisisk, spansk, tysk, svensk, dansk og norsk bokmål, alle med grunnlag i føringsgrammatikk. Flere av språkpara i den regelbaserte oversettelsesplattforma Apertium (http://wiki.apertium.org) bruker også føringsgrammatikk.
6. http://connexor.com tilbyr slike analyser for engelsk, fransk, spansk, tysk, svensk og finsk.
7. http://gtweb.uit.no/korp
8. http://sprakbanken.gu.se, 1.12. 2016.
9. Nunavut Hansard, parallellkorpus over det kanadiske parlamentet, http://www.assembly.nu.ca/hansard. En setningsparallellisert versjon er tilgjengelig på http://inuktitutcomputing.ca/NunavutHansard/ (1.12.2016).

Brown-korpuset var balansert til å speile sjangerfordelinga av publiserte tekster i USA i 1961, til sammenlikning har SIKOR langt mindre skjønnlitteratur (2 % mot 47,6 % i Brown), og langt mer administrativ tekst (32 % mot 6 % i Brown) og avistekst (50 % mot 17,9 % i Brown). Skjønnlitteratur er underrepresentert i SIKOR på grunn av uavklarte rettighetsspørsmål, men selv med full tilgang til all skjønnlitteratur ville andelen skjønnlitteratur sannsynligvis vært lavere for samisk enn for et majoritetsspråk.

Tabell 1. SIKOR-korpusets sju sjangere, antall ord (uten skilletegn) og deres prosentvis fordeling. Tallene inkluderer bare tekster som er skrevet med gjeldende ortografi (fra 1980).

Sjanger	**Antall ord**	**Prosentandel**
Avistekster	12.503.401	50 %
Administrative tekster	8.312.471	32 %
Faktatekster	2.019.958	8 %
Vitenskapelige tekster	954.744	4 %
Juridiske tekster	529.270	2 %
Skjønnlitterære tekster	479.371	2 %
Religiøse tekster	292.533	1 %
Ialt	25.091.749	100 %

Målt i antall sider[10] ville SIKOR utgjøre i overkant av 104.500 sider. Det tar i overkant av 3 timer[11] å gi korpuset en automatisk analyse av morfologi og syntaks[12] som f.eks. gjør det enkelt å lage frekvenslister for lemmaer, inkludert alle bøyingsformene, og for grammatiske konstruksjoner.[13]

Hele korpuset blir analysert, også setninger som inneholder ord som ikke blir konvertert slik at de blir leselige, eller har ord som ikke gjenkjennes av analyseprogrammet. 0,38 % av ordene i korpuset blir ikke gjenkjent av Giella-sme

10. Sidetall er målt som om det skulle være skrevet ut på A4 ark, med 2000 tegn pluss blanktegn på hver side.
11. Selve analysen tar 28 timer og 21 minutter, men korpuset blir analysert på en server med 16 parallelle prosesser, https://www.notur.no/hardware/stallo.
12. Denne artikkelen bygger på analysen som blei gjort 30.10.2016.
13. Frekvenslister for lemmaer og ordformer er tilgjengelige på http://giellatekno.uit.no/lex.en.html

fordi de er skrevet med versaler (de fleste akronymer blir gjenkjent). Det å analysere den lille delen av ord som er skrevet med versaler ville ha ført til flere ulemper i andre deler av analysestrengen uten at det hadde bidratt med ny lingvistisk innsikt, så vi har så langt ikke prioritert å løse dette problemet.

I korpuset blir så 3,2 % av de resterende ordene analysert som 'ikke gjenkjent'. Hele 29,0 % av disse ukjente ordene blir gjenkjent av analyseprogrammet for norsk bokmål, og det meste av dette er norske sitater i den samiske teksten. Det vil si at analyseprogrammet har dekningsgrad for 97,8 % av korpuset, når vi holder utenom ord skrevet med versaler og norske sitater.

Da gjenstår det 2,2 % som 'ikke gjenkjente' ord. Mange av disse er danske ord (*Kundskab*), engelske ord (*Author*), egennavn (*Vogelius*) eller unormerte skriveformer som programmet ikke kjenner (*árvvoštalla* pro *árvvoštallá*) 'å vurdere, presens Sg3', selv om programmet kjenner igjen *boahta* pro *boahtá* 'å komme, presens Sg3' med samme type unormert skriving. Dette tyder på at denne typen unormerte skriveformer ikke er lagt til systematisk, men bare for enkeltord eller grupper av ord. Det er en del tastefeil (*Sámpi* pro *Sápmi* 'Sameland'), og til sist er mye resultat av feilkonvertering av originalfilene (f.eks. ukjente tegn, halve ord eller to ord uten mellomrom). Feilkonverteringa er hovedsakelig forårsaket av ulike løsninger for å skrive samiske bokstaver (særlig fra tida før og like etter introduksjonen av tegnsettstandarden Unicode), av konvertering av pdf-filer til tekst og av problemer med OCR-lesning. Slike problemer blir forsøkt løst ved å legge til regler i konverteringsfila for hvert enkelt dokument, uten å endre originaldokumentet.

4 Den morfologiske modellen

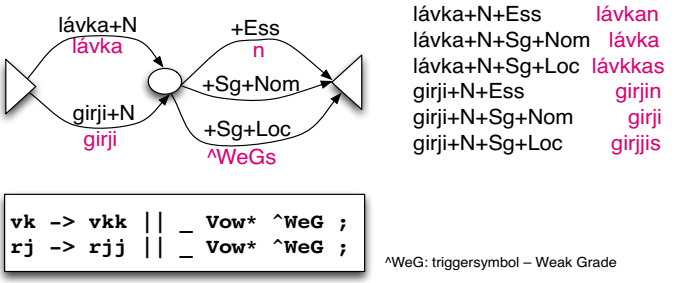
Dette kapitlet presenterer det teoretiske grunnlaget for den morfo(fono)logiske modellen for nordsamisk, ser på hvordan ordsammensetning og ordavledning blir modellert, og hva dette har å si for analysen av ord i korpuset sett fra et syntaktisk og et morfologisk perspektiv. Vi ser også nærmere på sammensetning som en produktiv morfologisk prosess.

4.1 Morfologi: To transdusere

For språk med mye morfologi og lite digitalt tekstkorpus, som på langt nær dekker alle ordformene i språket, er den beste løsning å lage endelige tilstandsautomater (se Antonsen og Trosterud (2010) for diskusjon om alternative metoder som stemming og statistisk tilnærming).

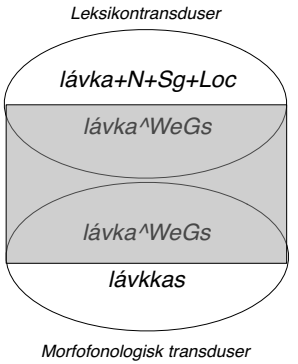
Automatene består av stier som lager alle mulige kombinasjoner av stammer og affikser i språket. Ved å legge suffikset *-n* til *lávka* 'ryggsekk, veske' og *girji* 'bok' får man ordformene *lávkan* og *girjin*. En grammatisk transduser (se Beesley og Karttunen 2003) er en type av automater hvor hver ordform har to representasjoner, ordforma og det korresponderende lemmaet + grammatiske tagger (det grammatiske ordet). Ordforma *lávkan* har også representasjonen *lávka+N+Ess*, og *girjin* har også representasjonen *girji+N+Ess*. Transduseren går begge veger og kan bli brukt både til analyse av ordforma og for å generere ordforma ved hjelp av lemma + grammatiske tagger. En slik transduser kalles en endelig tilstandstranduser, eller FST (eng. 'finite state transducer').

På grunn av de suprasegmentale morfofonologiske vekslingene i vokal- og konsonantsenteret har det nordsamiske analyseprogrammet to transdusere: en for leksikon og suffikser, og en for morfofonologiske prosesser (se Moshagen m.fl. (2004); Trosterud og Uibo (2005) for ei drøfting av sentrale trekk ved den morfofonologiske analysen av nordsamisk). I leksikontransduserens øvre nivå finnes det grammatiske ordet, mens nedre nivå utgjør en symbolstreng bestående av stamme og affikser, i tillegg til symboler som trigger morfofonologiske prosesser. Figur 1 og 2 (neste side) viser hvordan leksikontransduseren og den morfofonologiske transduseren arbeider sammen. Leksikontransduserens nedre nivå er innputt til den morfofonologiske transduseren, og sluttresultatet er ordforma. Morfofonologiske vekslinger gjøres også uten triggersymboler, og da på grunnlag av den fonologiske konteksten.



Figur 1. Illustrasjon av hvordan transduseren produserer ordformer av lemmaene *lávka* og *girji*. Symbolet *^WeG* legges til i lexc-transduseren og trigger endringer i konsonantsentrum: *vk:vkk* og *rj:rjj*.

Det er tilsammen 173.000 ord i leksikontransduseren, hvorav substantiv utgjør størstedelen: 90.000 fellesnavn og 49.000 egennavn, og av disse igjen er 17.000



Figur 2. Leksikontransduserens øvre nivå er det grammatiske ordet, mens det nedre nivået består av stamme, suffiks og symbolet *^WeG*. I den morfofonologiske transduseren er symbolet *^WeG* trigger for endringa *vk:vkk*. Sluttresultatet er ordforma *lávkkas* som dermed korresponderer med *lávka+N+Sg+Loc*. Det grå feltet forsvinner i sammensetninga av de to transduserne, og resultatet er en tovegs transduser for paret *lávka+N+Sg+Loc:lávkkas*.

nordsamiske navn og 31.500 er andre navn. Av de øvrige større ordklassene er 15.000 verb, 9200 adjektiv, og 4500 adverb.

Leksikontransduseren ble opprinnelig bygd opp med ordbøker av Pekka Sammallahti (Sammallahti 1989) og Nils Jernsletten (Jernsletten 1983) som basis, med i alt rundt 30000 lemmaer, men seinere er det blitt lagt til betydelige mengder fra korpustekster. For egennavn inneholder Giella-sme et ekstensivt leksikon av navn fra språkområdene som er relevante for samisk tekst. I og med at alle egennavn blir bøyd i kasus, og tradisjonelle samiske navn i tillegg har morfofonologiske vekslinger inne i navnet, har man på denne måten vært i stand til å identifisere den grammatiske funksjonen navnene har i setninga.

Arbeidet med å bygge en grammatisk modell for nordsamisk startet opp med et utkast til en morfofonologisk og morfologisk analyse laget av Pekka Sammallahti i 1993 (se Moshagen m. fl. 2004). Arbeidet ble tatt opp igjen i år 2000 (jf. Trosterud 2002), og videreført ved UiT. Den nordsamiske modellen som presenteres her er den første som ble gjort for et samisk språk, senere har det ved UiT blitt utarbeidet grammatiske modeller også for sør-, lule- og enaresamisk (jf. Antonsen og Trosterud 2010, 2011).

Den morfofonologiske transduseren består av 112 regler. Av disse styrer 46 regler stadieveksling i konsonantsenteret, som i figur 1, og 36 regler styrer endringer i stammevokalen, f.eks. *i:e* i forbindelse med sammensetning: *girji:girje-* 'bok'. 15 regler tar seg av endringer av stammekonsonant, f.eks. *žž:š* i *sápmelažžan:sápmelaš*, og 15 regler styrer diftongforenkling, som *uo:u* i *guolli:guliid* 'fisk'.

Ved å bare lage stier for normerte ordformer i transduseren, vil mange ord ikke få analyse. I transduseren er det også ikke-normerte former som finnes i korpus, f.eks. *ráhkanahttit* som er en ikke-normert form av *ráhkkanahttit* 'å forberede'.

Mange av de ikke-normerte formene som legges til transduseren, er alternative analyser av normerte ordformer. For eksempel vil forma for nominativ entall av *suohkan* 'kommune' også gi analyse som ikke-normert variant av genitiv og akkusativ form. Dette gir transduseren potensiale til å gi bedre analyse når forma er brukt slik i setninger, men samtidig gjør dette det vanskeligere å velge den ene rette forma utfra konteksten, fordi det blir flere former å velge mellom — se del 5.1 om disambiguering — men for f.eks. maskinoversetting er det viktig at transduseren takler hele kildespråket, også ikke-normerte former og variasjon.

Ved å merke former i leksikon og stier som er utafor normen med spesielle tagger, f.eks. *+Err/Orth* (eng. 'error orthography'), kan disse filtreres bort for å lage en transduser for ordretteprogram, og det er også mulig å generere normativt korrekte paradigmer for språklæringsprogrammer og e-ordbøker.

Den nordsamiske transduseren har vært i fokus for nesten alt arbeid med samisk språkteknologi. Transduseren er tilpasset skriftspråket i tekster, og er ikke tilpasset transkribert talespråk, og SIKOR inneholder foreløpig ikke transkribert talespråk. Den gjenkjenner bare nyeste nordsamiske ortografi. Ortografier som var i bruk før 1980, Bergsland/Ruong-ortografien (i Norge og Sverige) og Itkonen (i Finland), blir ikke gjenkjent. Eldre tekster må skannes for å få dem i digital form, og korrekturleste versjoner er foreløpig ikke tilgjengelige. I framtida vil det være aktuelt å lage konverteringsrutiner fra eldre til gjeldende ortografi.

Den nordsamiske transduseren er en av de største åpent tilgjengelige språkmodellene, jf. tabell 2 (neste side), som inneholder en oversikt over noen større tilgjengelige transdusere, målt i antall lemma.

Språk	Lemma	Kilde		Språk	Lemma	Kilde
Finsk	923.766	Omorfi		Enaresamisk	44.830	Giella-smn
Svensk	507.800	Hfst		Kvensk	41.169	Giella-fkv
Engelsk	146.600	Hfst		Tyrkisk	37.131	Hfst
Nordsamisk	146.102	Giella-sme		Kasakhisk	33.846	Apertium
Færøysk	88.097	Giella-fao		Kirgisisk	17.842	Apertium
Sørsamisk	58.643	Giella-sma		Præriecree	16.685	Giella-crk
Lulesamisk	48.394	Giella-smj				

Tabell 2. Åpent tilgjengelige transdusere.[14]

4.2 Orddanning

De sentrale orddanningsprosessene i nordsamisk er ordavledning og sammensetning. Ordavledning er en sentral del av grammatikken, ting som i f.eks. norsk ville ha blitt uttrykt med hjelp av verb pluss adverb blir i nordsamisk uttrykt ved hjelp av avledete verb. Ingen grammatisk analyse av nordsamisk er dekkende uten å ha en uttømmende behandling av orddanning.

4.2.1 Ordavledning

Avledning behandles på samme måte som bøying, men fordi resultatet av en avledning er et nytt leksem, vil stien gjennom transduseren peke fra avledningssuffikset og til bøyingsmorfologien for det resulterende leksemet. Ulikestavelsesverb vil f.eks. ha en sti for å legge til *-eapmi* paret med taggen *+Der/NomAct* for å få handlingsnomen, slik som *čuovvul(it) + eapmi => čuovvuleapmi* (*čuovvuleapmi:čuovvulit+V+Der/NomAct*) 'å følge opp => oppfølging', og stien vil fortsette til et eksisterende leksikon som legger til suffikser og triggersymboler til denne typen substantiv. Dermed vil *čuovvuleapmi* få hele bøyingsparadigmet og avledninger for substantiv, og dette gir par som *čuovvuleapmái:čuovvulit+V+Der/NomAct+N+Sg+Ill*.

Svært mange avledninger er allerede oppført i leksikonet, dvs. de er leksikalisert, dette gjelder også *čuovvuleapmi*. I tillegg til å utgjøre base for

14. De ulike transduserne er tilgjengelig på disse adressene: Hfst: https://sourceforge.net/projects/hfst/files/, Omorfi: https://github.com/flammie/omorfi, Giella: https://victorio.uit.no/langtech/trunk/langs/, Apertium: https://svn.code.sf.net/p/ apertium/svn/languages/

avledning med *eapmi er čuovvulit* også en inkoativ avledning av bevegelsesverbet *čuovvut* (Nickel og Sammallahti 2011: 555), dermed gir transduseren tre analyser av ordforma *čuovvuleapmái*:

čuovvuleapmi+N+Sg+Ill
čuovvulit+V+TV+Der/NomAct+N+Sg+Ill
čuovvut+V+TV+Der/l+V+Der/NomAct+N+Sg+Ill

Avledningstaggen *+Der/l* navngir den morfologiske prosessen (”legg til avledningssuffikset *–l-“)*, men ikke om den semantiske betydninga til prosessen, som varierer fra stamme til stamme. F.eks. vil den samme taggen brukes for avledning i tilfellet *borralit:borrat+V+TV+Der/l+V+Inf* som har subitiv betydning (Nickel og Sammallahti 2011: 544). Avledningsprosesser der den grammatiske funksjonen går fram av suffikset, får en tagg som viser funksjonen (som avledning til handlingsnomen, *+Der/NomAct*), mens avledningsprosesser med variabel grammatisk funksjon for tagger som representerer prosessen heller enn den resulterende semantiske effekten (som *+Der/l, +Der/h, …*).

Alt etter behov kan man velge hvilken analyse som er mest hensiktsmessig. Den syntaktiske delen av analyseprogrammet vil prioritere det mest avledete lemmaet, altså *čuovvuleapmi*, også til maskinoversetting, men der vil systemet prøve å oversette *čuovvulit* hvis lemmaet *čuovvuleapmi* ikke finnes i transferleksikonet, og subsidiært lemmaet *čuovvut*. Den elektroniske ordboka vil kunne bruke alle tre lemmaene, både for gjenkjenning av ordet, og for å gi bøyingsparadigme til brukeren. Vi vil referere til en analyse som gir leksem, ordklasse og morfosyntaktisk analyse som en analyse der ordet er sett *utafra*, dvs. fra et syntaktisk, ordeksternt perspektiv. For *čuovvuleapmái* er dette den første analysen, ordforma er illativ entall av substantivet ČUOVVULEAPMI. En analyse som gir en så uttømmende morfologisk analyse som mulig vil inkludere morfologiske prosesser som sammensetning og avledning, vi vil kalle det å se ordet *innafra*.

Hvis man skal telle antall avledninger i en tekst, vil man foretrekke å se ordet innafra, dvs. analyse nr. 3 ovafor. Men svært mange av ordene i leksikonene er avledninger. Av lemmaene i substantivleksikonet er for eksempel nesten 5671 leksikaliserte handlingsnomen (*+Der/NomAct*), og 217 av verblemmaene er leksikaliserte *+Der/l*-avledninger.

Men transduseren gir informasjon om bare de avledningene som er mest produktive, basert dels på språklig intuisjon, dels på funn i korpus. Transduseren gir for eksempel ingen informasjon om en avledning som *-alit*, alle verb på

-alit er leksikalisert. For følgende verb gir denne avledninga resiprok betydning (Nickel og Sammallahti 2011: 576):

oaidnalit+V+Inf 'å se (hverandre)'
náitalit+V+Inf 'å gifte seg (med hverandre)'

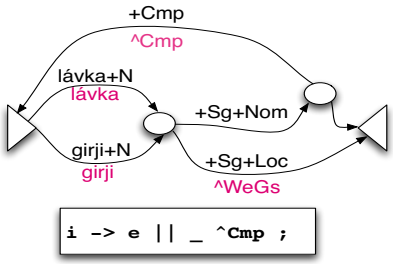
Valgene som gjøres ved bygginga av transduseren, får konsekvenser for hvilke programmer den kan brukes i. Store deler av arbeidet med det nordsamiske leksikonet har vært styrt fra behovene i ordretteprogrammet, og da er det viktigere å generere alle mulige ordformer, og unngå for mye overgenerering, dvs. unngå å bygge former som ikke finnes i språket, framfor å gi analyse hvor man ser ordet innafra. Dette vil si at for avledninger vil det ikke være mulig å få en full oversikt med transduseren slik den er bygd.

Noen avledningstyper er aldri leksikaliserte i leksikonet, slik som passivavledninga *-(oj)uvvot*, som i *čállit:čállojuvvot* 'å skrive'. Avledninga merkes i analysen med taggen *+Der/PassL* (= Passive Long, denne avledninga kalles av Nickel og Sammallahti (2011: 563) for intensjonelle passive verb). og inkoativforma *V-goahtit*, f.eks. *čállit:čálligoahtit* 'begynne å skrive' (merkes med taggen *+Der/InchL* = Inchoative Long, se Nickel og Sammallahti (2011: 554)). Heller ikke komparering av adjektiv leksikaliseres i leksikonet, og disse merkes ikke med *+Der*-tagg, men med *+Comp* og *+Superl* i analysen, for henholdsvis komparativ og superlativ.

4.2.2 Sammensetning

Nordsamisk har, som alle språka i det nordvesteuropeiske språkområdet, dynamisk sammensetning. I samisk dannes sammensetninger med førsteleddet fra mange ordklasser (se oversikt i Nickel og Sammallahti 2011: 662–671). Det er ikke praktisk mulig å leksikalisere alle nye sammensetninger som kommer inn i språket, og transduseren har da også mulighet for å lage dynamisk sammensetning. Det vil si at det for å modellere alle mulige substantiv-substantiv-sammensetninger går en sti fra hvert enkelt substantiv via en egen sammensetningsnode som legger til taggen *+Cmp*, og derifra tilbake til settet av substantiv igjen. Tilsvarende stier går fra adjektiv og verbets substantivavledninger til substantiv, fra en del adverb til substantiv, og fra numeraler til noen få adjektiv. Det går også stier fra substantiv til substantivavledninger av verb og adjektiv. Se tabell 3 for eksempler. Alle andre typer sammensetninger er leksikaliserte.

Med dynamisk sammensetning godkjennes også ord som ingen samisktalende av semantiske grunner ville bruke, og også uendelig lange sammensetninger, men de er vanligvis ikke problematiske fordi disse bare vil bli synlige hvis noen sender slike ord til analyse.



Figur 3. Det går en sti fra nominativforma tilbake til det første leksikonet. Stien inneholder taggen +*Cmp* kombinert med triggersymbolet ^*Cmp* som trigger vokalendring *i:e*. Transduseren gir formene *girjelávka* og *girjelávkkas* ('bokveske' i nominativ og lokativ), men også de semantisk mindre sannsynlige formene *lávkagirji* og *lávkagirjjis* ('veskebok'). Sammensatte former med forleddet i lokativ, **girjjislávka*, vil ikke bli gjenkjent.

Men for å kunne bruke transduseren til ordretteprogram og andre skrivestøtteverktøy, må det lages begrensninger for hvilken form sammensetningene kan ha. Hvis ikke noe annet er angitt, vil forleddet være i entall nominativ. I leksikonet er det lagt til en tagg +*CmpN/SgG* (*N* står her for *normativ*) for de substantivene som kan være forledd i genitiv entall. Dette gjelder ord som etterleddet på en eller annen måte hører inn under, eller er avkom og produkter av, som *gusamielki* 'kumelk'. Også for en del geografiske betegnelser er forleddet i genitiv, som *joganjálbmi* 'elvemunning', men dette varierer fra substantiv til substantiv (jf. eksempler i Nickel og Sammallahti 2011: 665–667). Et animat forledd kan være i flertall genitiv og merkes med taggen +*CmpN/PlG*, f.eks. *mánáidskuvla* 'barneskole'. Forleddets form kan også styres med en tagg +*CmpN-Left* i etterleddet. Her følger eksempler på tagger brukt for to ord.

(1) *loddi* 'fugl' har tagger for hvilke former ordet kan ha som forledd: +*CmpN/SgN* gir entall nominativ: *loddebivdu* 'fuglejakt', +*CmpN/SgG* gir genitiv entall: *lottečivga* 'fugleunge', +*CmpN/PlG* gir flertall genitiv: *lottiidráfáidahttin* 'fuglefredning')

(2) *lávlun* ‘synging’ har tagger for å styre forleddets former: +*CmpN/SgNomLeft* krever entall nominativ: *sálbmalávlun* ‘salmesang’, +*CmpN/SgGenLeft* krever genitiv entall: *sálmmalávlun* ‘salmesang’, +*CmpN/PlGenLeft* krever flertall genitiv: *mánáidlávlun* ‘barnesang’)

Hovedregelen er at forledd i nominativ entall skal ha allegroform (se Nickel og Sammallahti 2011: 22–23), det vil si at stammevokalen forkortes, som i *loddi* > *lodde-* i ordet *loddebivdu*, se eksempel (1). Forledd i genitiv entall skal etter hovedregelen ha largoform, dvs. uten vokalforkortning, *lotti-*. Men det finnes unntak fra disse to reglene på leksikalsk nivå, og noen forledd kan ha enten allegro- eller largoform. Dette løses i transduseren ved å dele substantivene i forskjellige leksikon og lage stier fra enten allegro- eller largoformen, eller fra begge formene (Moshagen m.fl. 2008).

Det er i noen tilfeller for ordretteprogrammet behov for å styre substantivenes posisjon i sammensetninga. Dette kan styres med taggen +CmpNP, hvor P står for posisjon.

+*CmpNP/None*: Lemmaet vil ikke kunne inngå i dynamiske sammensetning, som for *máŋggas* ‘mange (om mennesker)’

+*CmpNP/Last*: Lemmaet kan bare være etterledd, som for *gaskka* ‘ildstål’. Som forledd vil ordene bli leksikalisert, Dette for å unngå uønskede sammensetninger med *gaskka* istedenfor *gaska,* som ville ha dekket over skrivefeil som ordretteprogrammet ikke bør analysere som korrekte ord, f.eks. **gaskkavahkku* pro *gaskavahkku* ‘onsdag’

+*CmpNP/First*: Lemmaet kan bare være forledd *4-čiegahas* ‘firkant’

Taggene som angir at forleddet er genitiv eller nominativ, og taggene som angir posisjon, gir bare begrensninger for bygginga av transduser for ordretteprogrammer, og ikke for analyse av korpus.

Selv om det ikke er mulig å legge til alle mulige sammensetninger i leksikon, har det likevel blitt gjort et forsøk på nettopp det: Av de 90.000 substantivstammene som utgjør substantivleksikonet i transduseren, er 70.000 merket som leksikaliserte sammensetninger, mens 20.000 er usammensatte substantiv. Grunnen til dette er dels at det skal være mulig å lage en stavekontroll med teknologier som ikke har dynamisk sammensetning, men også for stavekontroller med dynamisk sammensetning, vil dette være en fordel. Forslagsgeneratoren vil anse sammensatte ord som er leksikaliserte som mer sannsynlige enn dynamisk sammensetninger.

Transduseren er også brukt i digitale ordbøker, som ordboka *Neahttadigisánit*[15], som med hjelp av transduseren kan oversette ord selv om søkeordet er en bøyd form (Johnson m.fl. 2013). Hvis sammensetninger ikke er leksikaliserte i ordboka, blir de analysert som dynamiske sammensetninger med oversetting av forledd og etterledd. Neahttadigisánit inneholder 22.800 substantiv, av disse er 17.625 leksikalske sammensetninger. I en versjon av ordboka som fungerer uten internettilknytning er leksikalisering av sammensatte ord nødvendig å få oversettelse for disse, i og med at ei slik ordbok ikke vil ha tilgang til morfologisk analyse av det komplekse ordet via serveren ordboka ligger på (Antonsen m.fl. 2009).

Det arbeides med maskinoversetting mellom flere språkpar, bl.a. fra nordsamisk til sørsamisk, lulesamisk, finsk og bokmål (Antonsen m.fl. 2017). Maskinoversettingssystemet[16] klarer å oversette forledd og etterledd, men hvis ordene i kildespråket og målspråket ikke har samme sammensetningsmønster, må de leksikaliseres og settes i transferleksika.

De leksikalske sammensetningene dekker løpende tekst godt: I SIKOR-korpuset med 25 millioner løpende ord var 9,6 millioner ord substantiv, av disse var 350.000, eller 3,6 %, dynamiske sammensetninger, som ikke var dekket av de 70.000 leksikaliserte sammensetningene i leksikon.

4.2.3 Sammensetning som produktiv morfologisk prosess

I analysen av korpuset vil de leksikaliserte sammensetningene få to analyser: en analyse med det leksikaliserte ordet som lemma, og en morfologisk analyse av forledd og etterledd. Den syntaktiske delen av analyseprogrammet (se del 5) prioriterer den leksikaliserte analysen. For å analysere de morfologiske egenskapene ved tekst trengs det dermed en dobbelt analyse. Etter den syntaktiske analysen får de relevante ordformene en ny morfologisk analyse, som inkluderer ordinterne egenskaper, som avledning og sammensetning, uten hensyn til om de var leksikaliserte eller ikke, og da får 10,7 % av substantivene i korpuset dynamisk sammensetningsanalyse.

Men analysen av korpuset viser at ytterligere 0,6 % av substantivene (56.000 ordformer) i realiteten er sammensatte substantiv hvor både forledd og etterledd er ordformer som Giella-sme kjenner, selv om disse ikke får dynamisk sammensetningsanalyse, selv etter morfologisk analyse. Disse sammensetningene er derfor ikke med i analyser videre i dette kapitlet. Disse ordene for-

15. http://sanit.oahpa.no
16. http://jorgal.uit.no /

deler seg på ca. 600 forskjellige sammensatte substantiv (leksemer), og lista domineres av ord med adverb som førsteledd (f.eks. *ovttasbargu* ('sammenarbeid = samarbeid'). Noen ord har pronomen, vanligst er *ieš*, som førsteledd, f.eks. *iešmearrideapmi* ('selvbestemmelse'), noen er ordenstall, som *nubbigiella* ('andrespråk) og tallord, *ovttaidlohku* ('entall'). Her er også en håndfull forledd som er substantiv i illativ, som *barguimáhcaheapmi ('til-arbeid-tilbakeføring'*), lokativ *ortnegisdoallan* ('i-orden-holding = vedlikehold'), essiv som *buorringeavaheapmi* ('som-gode-bruk = utnyttelse') og komparerte adjektiv, som *unnimusstandárda* ('minstestandard').

Førsteledd	**Prosentandel**	**Eksempel**
Nominativ entall	68 %	*skuvlahistorjá* 'skolehistorie'
Nominativ entall: Verb handlingsnomen	9 %	*jorgalanbargu* 'oversettingsarbeid'
Genitiv entall	18 %	*sámegiella* 'samisk språk'
Genitiv flertall	2 %	*nuoraidossodat* 'ungdommers avdeling = ungdomsavdeling'
Adjektiv attributtform	2 %	*oktasaščoahkkin* 'fellesmøte'
Andre typer	1 %	*1700-lohku* '1700-tallet'
Sammenlagt	100 %	

Tabell 3. Fordelinga mellom forskjellige sammensetningstyper for ordsammensetninger med ett forledd. N=2.200.768 (settet av substantiv i SIKOR med ett førsteledd og ett etterledd).

Tabell 3 viser fordelinga mellom forskjellige sammensetningstyper for fellesnavn som får dynamisk analyse. Vi har bare tatt med sammensetninger med bare ett førsteledd og etterledd, det vil si 84 % av alle substantivsammensetningene i SIKOR. Ved homonymi mellom genitiv og nominativ, som i *suohkanbáhppa* 'sokneprest', er sammensetninga telt bare som nominativ. Verbets handlingsnomen har som førsteledd i nominativ entall en egen kortform, *jorgalan* istedenfor den fulle forma som er *jorgaleapmi* (av *jorgalit* 'å oversette'). Sammensetninger med adjektiv som førsteledd kan ha samme typer bøyingsformer som substantiv, og i denne tabellen blir disse slått sammen. Forledd med adjektiv i attributtform regnes for seg. Andre typer inneholder sammensetninger med akronymer, forkortninger og tallord i førsteleddet. Tallene fra korpus viser at det er sammensetninger med førsteleddet i nominativ og genitiv entall som dominerer: de utgjør 95 % av alle sammensatte substantiv.

Vi vil her se nærmere på de ulike typene, se på hvilken sammensetningstype som er mest produktiv, av de dominerende typene, og også sammenligne produktiviteten med noen sentrale avledningsprosesser. For å sammenligne produktiviteten for de ulike prosessene tar vi utgangspunkt i Harald Baayen sine teorier om produktivitet. I følge Baayen (1993) kan produktivitet bli sett på som evnen til å produsere nye (usette) former. For en gitt morfologisk prosess kan produktiviteten dermed bli definert som tallet på unike eksemplar (hapax legomena, eller hapakser) delt på det totale antallet av eksemplar av prosessen.

$$P = n_1 / N$$

Tabell 4 viser produktivitet for sammensetning og for noen vanlige avledningsprosesser.

Type	n_1	N	$P = n_1/N$	n_1	$P = n_1/25113$
Sammensetning (lemma)	98.350	1.602.886	6,14 %	5987	23,82 %
Avledete handlingsnomen	5289	70759	7,47 %	1692	6,73 %
Inkoativ verbavledning	1692	25133	6,73 %	2128	8,47 %
Passivavledning - (oj)uvvot	3066	358.743	0,85 %	879	3,50 %

Tabell 4. For hver av de morfologiske prosessene gir n_1 = hapakser (unike eksemplar) , N = det totale antallet eksemplarer, og $P = n_1/N$ er et mål på produktiviteten. Hapaksene er målt av leksem, ikke av ordformer. De to siste kolonnene gir n_1 og P for et normalisert utvalg N = 25113 for alle fire orddannningstyper.

Slik produktivitet P er forstått her er produktivitet et graderbart begrep. For avledete handlingsnomen er det slik at for hvert nye eksemplar vil vi ha 7,47 % sjanse for å finne et eksemplar vi ikke har sett før. For inkoativ avledning (*bargagoahtit* "begynne å arbeide", fra *bargat* "arbeide") er det tilsvarende tallet 6,73 %, osv. P vil synke etter hvert som N øker, så for å normalisere P har vi også estimert P separat for N lik den minste av de tre kategoriene (her: inkoativ verbavledning). Det gir oss en høyere P-verdi for den mest frekvente prosessen, som er sammensetning.

Hvis vi ser nærmere på de tre ulike forleddstypene ved sammensetninger ser vi at de skiller seg fra hverandre. I korpuset er de ulike forleddstypene ulikt fordelt. Minst produktiv er genitiv entall som forledd (som i *sámegiella* "samisk språk"). Deretter kommer genitiv flertall (som i *mánáidgárdi* "barnehage"), og mest produktiv er nominativ entall (som i *sátnegirji* "ordbok"). Også her har vi normalisert for N, jf. tabell 5.

Forleddstype	**n_1**	**N**	**P = n_1/N**	**n_1**	**P = n_1/36143**
Nominativ entall	87.749	134.640	6,51 %	7686	21,27 %
Genitiv flertall	1794	36.143	4,96 %	1794	4,96 %
Genitiv entall	8850	218.103	4,06 %	1236	3,42 %

Tabell 5. Ulik produktivitet (P) for sammensetning med tre ulike forleddstyper. For hver type er n_1 = hapakser (unike eksemplar) , N = det totale antallet eksemplarer, og P = n_1/N et mål på produktiviteten. Hapaksene er målt av leksem, ikke av ordformer. De to siste kolonnene gir n_1 og P for et normalisert utvalg N = 36143 for alle tre forleddstyper.

De tre typene av forledd skiller seg fra hverandre også på andre måter. Punktlista nedenfor gir de 20 vanligste forleddene for hver grammatiske type, ordnet etter frekvens.

- **Nominativ entall**: *boazu* 'rein', *doaibma* 'virksomhet', *kultuvra* 'kultur', *giella* 'språk', *bargu* 'arbeid', *nisu/nisson* 'kvinne', *láhka* 'lov', *vuođđu* 'basis', *oahppu* 'utdanning', *álgu* 'begynnelse', *doarjja* 'støtte', *skuvla* 'skole', *álbmot* 'folk', *váldu* 'hoved-', *eana* 'jord', *joatkka* 'fortsettelse', *boahtti* 'kommende', *geassi* 'sommer', *oahpahus* 'undervisning', *stáhta* 'stat'
- **Genitiv flertall**: *mánná* 'barn', *oahppi* 'elev', *boanda* 'bonde', *nisu/nisson* 'kvinne', *buohcci* 'pasient', *olmmoš* 'menneske', *studeanta* 'student', *dálon* 'fastboende', *geavaheaddji* 'bruker', *ealli* 'dyr', *juovllat* 'jul', *guovlu* 'område', *dievdu* 'mann', *áhčči* 'far', *presideanta* 'president', *eadni* 'mor', *oahpaheaddji* 'lærer', *nieida* 'jente', *sápmi* 'same', *váhnen* 'forelder'
- **Genitiv entall**: *sápmi* 'same', *fylka* 'fylke', *luondu* 'natur', *ruoktu* 'hjem', *dárru* 'norsk', *riika* 'rike', *suopma* 'finsk språk', *guovlu* 'område', *boazu* 'rein', *gilli* 'bygd', *meahcci* 'utmark', *eadni* 'mor', *máilbmi*

‘verden’, *báhppa* ‘prest’, *ruotta* ‘svensk’, *dállu* ‘hus/gård’, *norga* ‘norsk’, *skuvla* ‘skole’, *leatna* ‘len’, *oapmi* ‘eiendel/husdyr’

For genitiv flertall har de tjue vanligste forleddene betydninga ”person” eller “dyr”, slik at etterleddet hører til eller er til for denne klassen av personer. Unntaket er *juovllat* som er et flertallssubstantiv. De tjue vanligste forleddene i genitiv entall har betydning ”område”, og etterleddet er en del av dette området, eller ”dyr”, med etterledd som en del av dyret, eller ”menneske”, og etterleddet er noe som er sterkt knyttet til vedkommende, som språk og familiemedlem. For *báhppa* er etterleddet oftest noe som administrativt er knyttet til prestestillinga. Der forleddet er i nominativ entall er det mer heterogent, det kan være både begrep, område og person som etterleddet assosieres med på en eller annen måte. Nominativ entall skiller seg med andre ord fra de to andre typene med å mangle restriksjoner både på forledd og på type av semantisk binding mellom for- og etterledd, og resultatet blir en forskjell i produktivitet, der nominativ entall er mest produktivt.

Forskjellen i produktivitet som vist i tabell 5 korresponderer også med frekvensprofilen for forleddene. Det mest brukte forleddet (henholdsvis *mánná* og *sápmi*) for genitiv flertall og entall utgjør hele 56.1% og 55.7% av den totale mengden av sammensetninger, mens det tilsvarende tallet for nominativ entall (*boazu*) er 3.3%.

5 Morfologisk analyse i kontekst

Vi presenterer her den syntaktiske modellen som velger riktig morfologisk analyse basert på konteksten ordforma står i. Den morfologiske og den syntaktiske modellen er bygd opp langs svært ulike prinsipper, noe som korresponderer med et modulært syn på grammatisk struktur.

5.1 Disambiguering

De samiske grammatikkene består av handskrevne regler, se under figur 4 for eksempler på regler. Regelformalismen er nærmere beskrevet i Trosterud og Wiechetek (2007).

I nordsamisk tekst har hver ordform i gjennomsnitt 2,6 mulige morfologiske analyser (Trosterud og Wiechetek 2007). For substantiv er det full homonymi mellom akkusativ og genitiv, og hvis substantivet ikke har stadieveksling, blir entallsformene homonyme også med nominativforma. Ulikestavelsesverb har flere homonyme former enn likestavelsesverb, og verbforma

leat 'å være' har fem forskjellige analyser: infinitiv, nektelsesform, og tre presensformer i indikativ: 2. person entall, 1. person flertall og 3. person flertall.

Den syntaktiske analysatoren velger analysen som er riktig utfra konteksten, dvs. at den fjerner ambiguiteten – den *disambiguerer*. Denne disambigueringa gjøres med manuelt skrevne kontekstsensitive regler, for nordsamisk er det 1835 regler som enten velger eller fjerner analyser og 202 regler som legger til eller endrer tagger, som hjelp i disambigueringa. Grammatisk er det mulig å arbeide på to måter, nedenfra og opp, eller ovenfra og ned[17]. Den siste typen prøver ut hypoteser om hva slags syntaktisk struktur det er i setningen, og dette er grunnlaget for disambiguering. Giella-sme arbeider nedenfra og opp, og dermed klarer den å gi analyse også til setningsfragmenter og komplekse setninger.

```
"<Mii>"
    "mun" Pron Pers Pl1 Nom
    "mii" Pron Indef Sg Nom
    "mii" Pron Interr Sg Nom
    "mii" Pron Rel Sg Nom
"<eat>"
    "ii" V IV Neg Ind Pl1
"<leat>"
    "leat" V IV Ind Prs ConNeg
    "leat" V IV Ind Prs Pl3
    "leat" V IV Ind Prs Sg2
    "leat" V IV Inf
    "leat" V IV Ind Prs Pl1
"<dan>"
    "dat" Pron Dem Sg Acc
    "dat" Pron Dem Sg Gen
"<muitalan>"
    "muitalit" V TV PrfPrc
    "muitalit" V TV Actio Nom
    "muitalit" V TV Actio Gen
    "muitalit" V TV Ind Prt ConNeg
    "muitalit" V TV Ind Prs Sg1
"<.>"
    "." CLB
```

```
"<Mii>"
    "mun" Pron Pers Pl1 Nom
"<eat>"
    "ii" V IV Neg Ind Pl1
"<leat>"
    "leat" V IV Ind Prs ConNeg
"<dan>"
    "dat" Pron Dem Sg Acc
"<muitalan>"
    "muitalit" V TV PrfPrc
"<.>"
    "." CLB
```

Figur 4. De fleste ordene i setninga *Mii eat leat dan muitalan* ('Vi har ikke fortalt det') får flere morfologiske analyser, slik som i analysen til venstre. Giella-sme bruker lingvistiske betingelser (føringer) i konteksten, for valg av riktig form, og analysen blir ideelt sett uten ambiguitet, som vist til høyre i figuren.

Figur 4 viser analyse av en setning før og etter disambiguering. Disambigueringa av denne setninga gjøres med følgende føringer:

17. Ovenfra-og-ned-innfallsvinkelen går tilbake til regelformat av typen *S -> NP VP* i klassisk generativ grammatikk, og er i nyere datalingvistikk representert av formalismer som for eksempel LFG og HPSG. For ei kort innføring i HPSG som prosesseringsformalisme, med referanser også til LFG, se Levine og Meurers (2006).

- For ordforma *mii* velges analysen Pron Pl1 fordi ordet etterfølges av V Pl1.
- For ordforma *leat* velges ConNeg fordi ordet etterfølger V Neg.
- For ordforma *dan* velges Acc fordi den etterfølges av et transitivt verb, og det ikke finnes noe annet ord med analysen Acc i mulig posisjon som objekt. Ordet står heller ikke etter Pr eller Num, eller foran Po, dvs. at den ikke er del av en pre- eller postposisjonsuttrykk, eller en numeralfrase.
- For ordforma *muitalan* velges PrfPrc fordi den er komplement til ConNeg.

De to største homonymiproblemene for substantiv har vært å disambiguere mellom komitativ entall og lokativ flertall, og mellom akkusativ og genitiv. Den førstnevnte utfordringen er på langt nær løst. Det er mulig å disambiguere komitativ entall vs. lokativ flertall ved å utnytte forskjellen i numerus: reglene viser til mulige modifikatorer som forteller at substantivet er i entall (f.eks. *dáinna* 'med denne') eller flertall, (f.eks. *máŋggalágan* 'mange slags'). Det er også mulig å la kombinasjoner av visse verb og substantiv gi komitativ tolkning, som i eksemplene (3-6). Likevel er ikke disse reglene uttømmende, og hvis ingen regel velger komitativ entall, blir ordet tolket som lokativ.

(3) *sii váldet oktavuođa skuvllain* 'de tar kontakt med skolen'
(4) *oahpaheaddjit barget skuvllain ovttas* 'lærerne arbeider med skolen sammen'
(5) *tiibmoaahpaheaddjit barget maiddái eará skuvllain* 'timelærerne arbeider også på andre skoler'
(6) *Rievvárat báhtaredje bolesiin.* 'Røverne flyktet fra politimennene' eller 'med politimannen'

I eksemplene (3–6) har ordformene *skuvllain* og *bolesiin* analyse både som komitativ entall og lokativ flertall. I eksempel (3) vil analysatoren velge komitativ på grunn av at den inneholder en regel om at verbkonstruksjonen til *váldet oktavuođa* (å ta kontakt) vil være komitativ når substantivet er en institusjon, som *skuvla*. I de to neste eksemplene er verbet *'bargat'* (arbeide), og i eksempel (4) vil kombinasjonen med adverbet *ovttas* gjøre at analysatoren velger komitativ (med skolen), men i (5) er det ingen ord som peker mot komitativ, og resultatet er da lokativ (på skolene). Eksempel (6) er tvetydig også for en

menneskelig leser, som på grunnlag av erfaring med røvere og politi, vil gå ut fra at lokativtolkninga er den riktige.

Det er full homonymi mellom akkusativ og genitiv, med unntak av noen tallord og ett pronomen i entall, og det er vanlig å kalle forma akkusativ-genitiv (f.eks. i Nickel og Sammallahti 2011). Men for regelbasert maskinoversetting er det en fordel å ha dette disambiguert, siden både lulesamisk og sørsamisk skiller mellom disse to kasusene. Også ved maskinoversetting fra nordsamisk til norsk, gir disambiguering mellom disse to kasusene det enklere å få en riktig oversetting.

I tillegg til å referere til setningas syntaktiske struktur, har en del ambiguitet blitt løst ved å legge til den semantiske taggen +*Sem/Hum* til substantiv som annoterer mennesker. I eksempel (7) vil man ved å referere til en slik tagg kunne lage en regel for *eatni* ('mor') som eier av kjøkkenet og dermed genitiv, i motsetning til eksempel (8) hvor *mállása* ('suppe, varm mat') ikke vil få denne taggen, og dermed vil analysatoren velge akkusativ som objektet for det transitive verbet *borrat* 'å spise'. Eksempel (9) derimot har en flertydighet som er vanskelig å løse uten en større kontekst. Objektet kan være *Biera* (et mannsnavn), som jeg ser *bak bilens dør*, eller det kan være *Bieras bil* som jeg ser *bak døra*. Hvis man bytter ut *uvssa* 'dør' med *viesu* 'hus', som i eksempel (10), blir det klart at det er *Bieras bil* som er objektet.

(7) *Mun boran eatni gievkkanis.* 'Jeg spiser på mors kjøkken.'
(8) *Mun boran mállása gievkkanis.* 'Jeg spiser suppe på kjøkkenet.'
(9) *Mun oainnán Biera biilla uvssa duohken*. 'Jeg ser [Bieras bil] [bak døra] eller [Biera] [bak bilens dør].'
(10) *Mun oainnán Biera biilla viesu duohken* 'Jeg ser Bieras bil bak huset.'

Reglene er bygd opp slik at det først er regler for spesielle tilfeller for å velge genitiv, deretter er det regler for å velge akkusativ og til sist blir alle gjenværende genitiv. Dette er en parallell til disambigueringa av komitativ og lokativ, og også mellom lokativ entall og substantiv i akkusativ eller genitiv med possessiv suffiks for 3. person entall, f.eks. *gielas* ('i språket' versus 'språket sitt'). Alternativet ville være å formulere regler for begge valgene og la resten stå med to analyser. En del programmer, som for eksempel maskinoversetting, kan bare ha en utputt, og må derfor velge en analyse.

5.2 Egenskaper ved verb og substantiv i korpuset – sjanger for sjanger

Disambigueringa er viktig for å vite hvilke ordklasser og bøyingsformer korpuset består av. Skrivefeil kan gi ord en annen bøyingsform enn det forfatteren hadde tenkt, og dette vil endre konteksten for ordene i setninga og kan dermed resultere i feil disambiguering for flere ord i setninga, selv om en del frekvente former er tatt høyde for i transduseren, slik at den syntaktiske analysen likevel blir riktig.

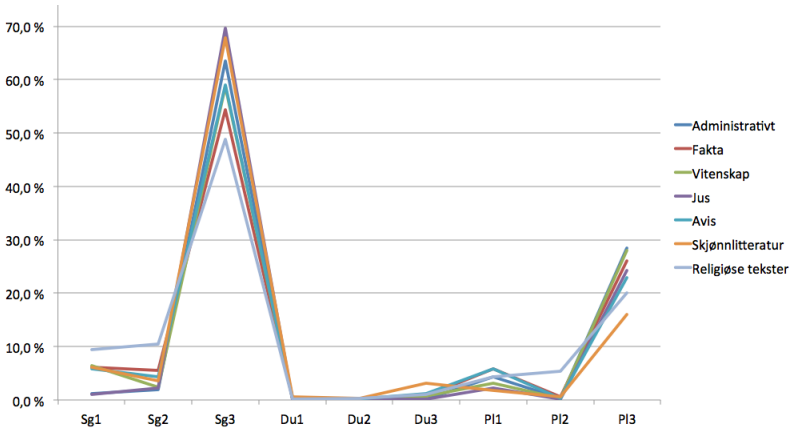
Ideelt sett skal hver ordform gi bare en analyse etter disambiguering, men ordene i analyse med hele korpuset har i gjennomsnitt 1,03 analyser. Men det er bare 2,7 % av ordene som har mer enn en analyse, og av disse er nesten en tredjedel akronymer, numeraler og forkortninger. For å få en syntaktisk analyse må alle argumenter tilordnes kasus. For akronymer og arabiske numeraler uttrykkes for eksempel ikke distinksjonen mellom nominativ og genitiv, og resultatet blir mer homonymi. Av de gjenværende ordene er den dominerende ambiguiteten ord uten stadieveksling som får både nominativ og genitiv analyse (33 %).

En tidligere evaluering av det nordsamiske analyseprogrammet viste at gjenkjenning av leksem + ordklasse var riktig i 99 % av tilfellene, og valg av korrekt full morfologisk analyse var riktig for 94 % av ordene (Antonsen m.fl. 2010). Denne evalueringa blei gjort på et lite korpus bestående av fulle setninger. SIKOR inneholder en del feilkonverterte ord og fragmenterte setninger, så disambiguering av hele korpuset vil nok gi et noe dårligere resultat enn den nevnte evalueringa. For å få best mulig kvalitet på analysen har vi for tallene videre i dette kapitlet fjernet alle setninger med minst ett ikke gjenkjent ord, bortsett fra for figur 6. Dette innebærer at grunnlaget for alle andre figurer enn figur 6 utgjøres av 76,7 % av korpuset (heretter referert til som 76,7-korpuset). Det kan da være relevant å nevne hvordan dette endrer sammensetninga av korpuset, i forhold til tallene presentert i tabell 1. Avistekster utgjør nå 49 %, administrative tekster 34 % og vitenskapelige tekster 3 %.

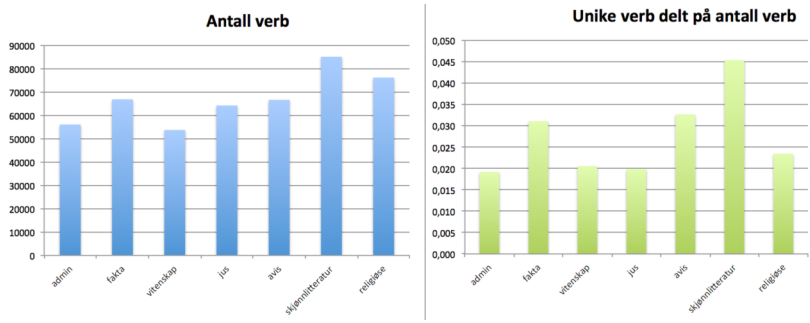
Verb i nordsamisk bøyes i person, tall (entall, totall, flertall), tempus og (fire) modus, i tillegg til at de har flere infinitte former. 23 % av ordene i korpuset er verb, og av disse er 64 % finitte verb.

Vi har latt analyseprogrammet beregne fordelinga av person-nummerbøying av finitte verb for hver sjanger i korpuset, se figur 5. Alle grafene har 3. person på topp, men variasjonen er likevel stor. Religiøse tekster har færrest former i 3. person, med 20 % Pl3 og 49 % Sg3, og disse tekstene ligger på topp for 1. og 2. person entall med henholdsvis 10 % og 11 %. Ikke overraskende

har juridiske tekster bare 1 % 1. person entall, og i skjønnlitterære tekster er det mer bruk av totallsformer enn i andre sjangre, sjøl om andelen er liten (i alt 4 %).



Figur 5. Fordelinga av finitte verb- former varierer mellom de forskjellige sjangerne i korpuset. Hver linje gir 100 % sammenlagt for alle verbformene. Data fra 76,7-korpuset.



Figur 6. Diagrammet til venstre er en sammenlikning av antall verb i hver sjanger. For å kunne sammenlikne antall verb, har vi for denne tabellen brukt alle de religiøse tekstene, som er den minste sjangeren, og så tatt et tilfeldig sammensatt delkorpus av samme størrelse fra de andre sjangerne (hvert på 292.000 ord). I diagrammet til høyre er antall unike verb delt på antall verb i delkorpusene.

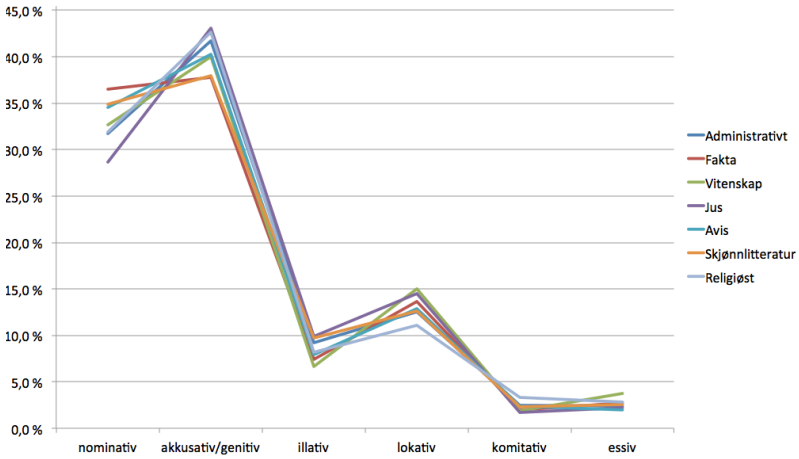
Analysen av korpuset viser at fordelinga mellom de forskjellige finitte formene ikke er så dramatisk forskjellig mellom de ulike tekstsjangerne. Men verbfrekvensen er svært forskjellig. Figur 6 (til venstre) viser at setningene i skjønnlitteraturen er atskillig rikere på verb enn de andre sjangerne, i denne sjangeren brukes 59 % flere verb enn i vitenskap, som har færrest verb. Religiøse tekster ligger nest høyest. Diagrammet til høyre viser ordforrådet for verb ved å dele antall unike verb på alle verb i delkorpuset, og skjønnlitteratur har forholdsvis nesten 2,5 ganger flere ulike verb enn administrative tekster, selv om man tar høyde for at det er flere verb sammenlagt. Her er det avistekster som ligger nest høyest, med faktatekster nesten på samme nivå.

Tempus	**Administrativt**	**Fakta**	**Vitenskap**	**Jus**	**Avis**	**Skjønnlitteratur**	**Religiøse tekster**
Presens	89 %	68 %	77 %	86 %	81 %	33 %	63 %
Preteritum	11 %	32 %	23 %	14 %	19 %	67 %	37 %
Tilsammen	100 %	100 %	100 %	100 %	100 %	100 %	100 %

Tabell 6. Tempusfordelinga av finitte verb i SIKOR, data fra 76,7-korpuset.

Også for tempus av verb er det store forskjeller fra sjanger til sjanger, se tabell 6. Sjangrene faller grovt sett i tre grupper. De fleste sjangrene ligger på 81–89 % for presens. Faktatekstene ligger noe under de andre, på 68 %, her er en stor del av tekstene hentet fra Samisk skolehistorie, et verk dominert av beskrivelser av historiske hendelser. Skjønnlitteratur er den eneste sjangeren med mindre presens enn preteritum, bare 33 %. Religiøse tekster kommer i en mellomposisjon, og disse tekstene svinger sjangermessig mellom skjønnlitteratur og saksprosa.

I utgangspunktet er det forventet at presens, som den umarkerte tempuskategorien (presens representerer også futurum i samisk), skal være klart sterkere representert enn det markerte preteritum. Dette stemmer da også, bortsett fra for skjønnlitteratur. For dette grammatiske trekket er altså skjønnlitteratur den minst representative sjangeren. Tydeligvis er det mange forfattere som bruker et fortellende modus der preteritum går igjen gjennom hele fortellinga, dette er en faktor å ta med i betraktning når en vurderer korpussjanger og representativitet. En annen mulig feilkilde er det svært avgrensede utvalget av skjønnlitteratur (bare 2% av korpuset), så skjevhet i utvalget kan også ha bidratt til avviket for skjønnlitteratur.



Figur 7. Fordelinga mellom forskjellige kasus for substantiv, med en graf for hver sjanger i SIKOR. Akkusativ og genitiv er sett under ett. Hver linje viser en del av korpuset, og kasusfordelinga langs hver linje er sammenlagt 100 %, data fra 76,7-korpuset.

I nordsamisk er det sju kasus for substantiv, men med full homonymi mellom akkusativ og genitiv, både i entall og flertall. Essiv skiller ikke mellom numerus. Vi har testet fordelinga av substantiv på forskjellige kasus, men ikke tatt med egennavn, akronymer eller forkortninger. Akkusativ og genitiv er slått sammen for å unngå en mulig feilkilde i disambigueringa.

Forskjellen mellom kasusfordelinga er størst for nominativ, hvor faktatekster har størst andel (37 %) og juridiske tekster minst andel (29 %). Disse to sjangrene plasserer seg motsatt når det gjelder akkusativ/genitiv, med juridiske tekster på topp (43 %) og faktatekster og skjønnlitterære tekster på bunn (38 %). Vitenskapelige tekster har forholdsvis større andel essiv (4 %) enn de andre sjangrene. Usikkerhet ved analyseresultene diskuteres etter tabell 7.

Tabell 7 viser kasusfordelinga for substantiv i hele korpuset, for entall og flertall. Essiv skiller ikke mellom numerus, og er tatt med under entall. Ordet *lassin* 'i tillegg, dessuten' får analysen som essivforma av substantivet *lassi*, og utgjorde nesten 10 % av essivformene. I og med at ordet står i kontekster der ingen andre substantiv i essiv er belagt i korpus, (*dasa lassin* 'i tillegg til det', men *dasa *vuostálasvuohtan / *kontrástan* "i motsetning/kontrast til det", ...) har vi sett dette som et leksikalisert adverb, og det er derfor ikke med i tallet

Tabell 7. Kasusfordelinga for substantiv og numerus i SIKOR, data fra 76,7-korpuset.

Kasus	Singular	Plural	Kasus i prosent av alle kasus
Nominativ	33,0 %	34,3 %	33,4 %
Akkusativ/genitiv	40,7 %	40,7 %	40,7 %
Illativ	7,9 %	9,9 %	8,4 %
Lokativ	13,2 %	11,9 %	12,9 %
Komitativ	2,0 %	3,1 %	2,3 %
Essiv	3,2 %	-	2,3 %
Tilsammen	100 %	100 %	100 %

for essiv i tabellen. Sammenlagt utgjør flertallsformene 29 % av substantivene, når man ser bort fra essiv.

En mulig feilkilde for tallene i tabellen er at Giella-sme må disambiguere homonymien mellom lokativ flertall og komitativ entall, som to komplementære analyser i korpuset. Vi plukket ut tilfeldige setninger fra korpus og evaluerte disambigueringa av disse to analysene for 400 ord[18]. Denne evalueringa overført til korpuset som helhet gir som resultat at prosenten for komitativ entall bør kunne økes fra 2,0 % til 2,2 %, og lokativ flertall bør kunne senkes fra 12,9 % til 12,2 %. Dette får også konsekvenser for grafene i figur 7, men forholdsvis mindre fordi entall og flertall her sees under ett.

Når vi sammenlikner resultatene som er gjort for 76,7-korpuset med resultatene for hele korpuset, finner vi den største forskjellen i verbprofilen, hvor andelen 2. persons bøyinger sammenlagt er 4,3 % istedenfor 3,0 % for hele korpuset. For kasusfordelinga er forskjellen 0–0,3 prosentpoeng for hvert kasus. Det er overraskende liten forskjell mellom resultatene. Årsaken er at i lange setninger vil et ord uten analyse vanligvis bare berøre disambigueringa for de nærmeste leddene. Dette viser at en analyse som bygges nedenfor og opp, fra ordform via lokal dependens og til full setningsanalyse, er en robust måte å bygge syntaktisk analyse på.

Kasusprofilen er overraskende lik fra sjanger til sjanger. På de leksikalske nivået er det derimot store forskjeller, noe som kommer fram i frekvenslister for substantiv. To substantiv, *sápmi* ‘same’ og *olmmoš* ‘menneske’ finnes blant

18. 42 av 260 ord med lokativanalyse var komitativ og 25 av 140 ord med komitativanalyse var lokativ.

de mest frekvente 10 substantivene for fem av sjangerne, og listene for administrative tekster og avistekster har flere sammenfall. Listene fra de juridiske og de religiøse tekstene er klart sjangerspesifikke, med ord som henholdsvis *láhka, mearrádus, vuoigatvuohta* 'lov, bestemmelse, rettighet' versus *Ipmil, Hearrá, bárdni, áhčči* 'Gud, Herren, sønn, far'. Lista for skjønnlitteratur skiller seg ut ved at blant annet ord for kroppsdeler er blant de ti mest frekvente.

6 Sammendrag og konklusjon

Vi har i denne artikkelen presentert Giella-sme, et analyseprogram for nordsamisk løpende tekst, bestående av en morfologisk transduser og en føringsgrammatikk for disambiguering og syntaktisk analyse. Giella-sme er en av de største åpent tilgjengelige transduserne for naturlig språk. Den er i bruk i et vidt spekter av applikasjoner, både vitenskapelige og praktiske.

Formelt er Giella-sme bygd opp som komposisjonen av en morfologisk og en morfofonologisk transduser, der den resulterende transduseren veksler mellom lemma + analyse på den ene sida og ordform på den andre. Et stort antall sammensetninger er leksikalisert, og lagt til transduseren som sådan, men dette er ikke nok til å analysere løpende tekst, og transduseren analyserer også dynamisk sammensetning og avledning.

Føringsgrammatikken gir en kontekstuell disambiguering av morfologisk analyse, med en høy presisjon (99 %) på grunnleggende analyse (lemma + ordklasse). Transduseren har en dekningsgrad på 97,8 % i det nordsamiske korpuset SIKOR, når vi holder utenom ord skrevet med versaler og norske sitater. En stor del av de resterende 2,2 % er tilfeldige ord på dansk eller engelsk, det er egennavn eller unormerte skriveformer som transduseren ikke kjenner igjen. Datamaskinell prosessering av et språk med så kompleks morfologi som nordsamisk vil alltid kreve lemmatisering og grammatisk analyse, analyseprogrammet presentert her gjør en slik prosessering mulig. Hele det samiske analyseprogrammet er tilgjengelig både som åpen kildekode og som en onlinetjeneste.[19]

Giella-sme vil i forskjellig grad kunne se ordene innafra og utafra, dvs. analysere dels sammensetnings- og avledningsstruktur, og dels morfosyntaktiske trekk, hvordan ordet er bøyd. Den syntaktiske analysatoren velger den mest leksikaliserte analysen av hver ordform. For å se nærmere på den interne strukturen av ordet må vi først bruke den syntaktiske analysatoren til å

19. http://giellatekno.uit.no/cgi/index.sme.nob.html

plukke ut relevante ord (for eksempel alle substantiv), og deretter må dette utvalget analyseres en ekstra gang bare med den morfologiske analysatoren.

Vi så nærmere på produktiviteten til en del sentrale orddanningsprosesser for substantiv og verb. Nordsamisk har mange avledningsprosesser og dynamisk sammensetning. Med å måle fordelinga av hapax legomena over totalt antall avlednings- og derivasjonstyper, har vi vist at sammensetning er langt mer produktiv enn avledning. Kategorien passiv, som er en avledningskategori i nordsamisk, er bare marginalt produktiv, sammenlignet med inkoativ. For sammensetning substantiv + substantiv er det mulig med tre ulike forleddstyper, nominativ og genitiv entall, og genitiv flertall. Her har de siste to typene restriksjoner både for type forledd og semantisk binding, og de er da også langt mindre produktive enn nominativ entall, som ikke har slike bindinger.

Halvparten av tekstene i SIKOR er avistekster, og de skjønnlitterære tekstene utgjør bare 2 %. For visse sider ved samisk grammatikk er denne fordelinga av korpuset til ulempe, for andre sider ser det ut til at det er liten variasjon sjangrene i mellom. Kasus er en kategori med liten variasjon, dette er også en grammatisk kategori som i liten grad peker ut over setninga. I skarp kontrast står særlig tempus og modus for verb, der for eksempel preteritum varierer mellom 11 % og 67 % fra sjanger til sjanger. Vi hadde også forventa stor variasjon i person og numerus for verb, og fant det til en viss grad. Juridiske tekster har sterk overvekt for tredje person entall (70 %) og nærmest fravær av første person, mens religiøs tekst har i underkant av 10 % første person entall og bare 43 % tredje person entall.

Kilder

SIKOR. UiT Norges arktiske universitet og Sametinget i Norges samiske tekstkorpus. Versjon 08.12.2016, URL: http://gtweb.uit.no/korp/

Referanser

Antonsen, Lene, Ciprian Gerstenberger, Sjur Moshagen & Trond Trosterud. 2009. Ei intelligent elektronisk ordbok for samisk. *LexicoNordica* 16. Oslo: Nordisk forening for leksikografi, 271–283.

Antonsen, Lene, Linda Wiechetek & Trond Trosterud. 2010. Reusing Grammatical Resources for New Languages. I *Proceedings of the International conference on Language Resources and Evaluation LREC*

2010. ISBN 2-9517408-6-7. Stroudsburg: The Association for Computational Linguistics, 2782–2789.

Antonsen, Lene, Saara Huhmarniemi & Trond Trosterud. 2009. Constraint Grammar in Dialogue Systems. Workshop: Constraint grammar and robust parsing.I *NEALT Proceedings Series 2009* Volum 8, 13–21.

Antonsen, Lene & Trond Trosterud. 2010. Manne dihtor galgá máhttit grammatihka? (Why the computer should know its Sami grammar.) *Sámi Dieđalaš Áigečála* 1/2010, 3–28.

Antonsen, Lene & Trond Trosterud. 2011. Next to nothing – a cheap South Saami disambiguator. Workshop: Constraint Grammar Applications. I *NEALT Proceedings Series 2011* Volum 14 [10], 1–7.

Antonsen, Lene, Ciprian Gerstenberger, Maja Kappfjell, Sandra Nystø Rahka, Marja-Liisa Olthuis, Trond Trosterud & Francis M. Tyers 2017: Machine translation with North Saami as a pivot language. *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017, Gothenburg, Sweden.* NEALT Proceedings Series. Linköping University Electronic Press, Linköpings universitet. p. 123–131.

Antonsen, Lene & Trond Trosterud. 2017. *A computational model of the Inari Saami morphophonology.* Manuskript.

Arppe, Antti. 2005. The Very Long Way from Basic Linguistic Research to Commercially Successful Language Technology: the Case of Two-Level Morphology. I Arppe, Antti, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund & Anssi Yli-Jyrä (red.): *Inquiries into Words, Constraints, and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday.* CSLI Studies in Computational Linguistics ONLINE, 2–17.

Baayen, Harald. 1993. Quantitative aspects of morphological productivity. I Booij, G. E. & J. van Marle (red.): *Yearbook of Morphology 1991*. Dordrecht: Kluwer Academic Publishers, 109–149.

Beesley, Kenneth R. & Lauri Karttunen. 2003. *Finite State Morphology.* Stanford, California: CSLI publications in Computational Linguistics.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Huhmarniemi, Sara, Sjur N. Moshagen & Trond Trosterud. 2007. Usage of XSL Stylesheets for the annotation of the Sámi language corpora. I *LAW '07: Proceedings of the Linguistic Annotation Workshop*. Morristown, NJ, USA: Association for Computational Linguistics, 45–48.

Jernsletten, Nils 1983: *Álgosátnegirji : samisk-norsk ordbok*. Oslo : Universitetsforlaget.

Johannessen, Janne Bondi, Kristin Hagen, André Lynum & Anders Nøklestad. 2012. OBT+stat. A combined rule-based and statistical tagger. I Andersen, Gisle (red.): *Exploring Newspaper Language. Corpus compilation and research based on the Norwegian Newspaper Corpus*. John Benjamins Publishing Company, 51-65.

Johnson, Ryan, Lene Antonsen & Trond Trosterud. 2013. Using finite state transducers for making efficient reading comprehension dictionaries. I *Proceedings of the 19th Nordic Conference of Computational Linguistics NODALIDA 2013* (NEALT Proceedings Series 16), 59–71.

Karlsson, Fred, Juha Heikkilä & Arto Anttila. 1995. *Constraint grammar. A language-independent system for parsing unrestricted text.* Berlin – New York: Mouton de Gruyter.

Karlsson, Fred. 1990. Constraint grammar as a framework for parsing running text. I Hans Karlgren (red.): *COLING -90: Papers Presented to the 13th International Conference on Computational Linguistics on the Occasion of the 25th Anniversary of COLING and the 350th Anniversary of Helsinki University*. Vol. 3. Helsinki: Yliopistopaino, 168–173.

Karttunen, Lauri. 2000. Applications of Finite-State Transducers in Natural Language Processing. I *Revised Papers from the 5th International Conference on Implementation and Application of Automata*. London: Springer-Verlag, 34–56.

Karttunen Lauri & Kenneth R. Beesley. 2005. Twenty-Five Years of Finite-State Morphology. I Antti Arppe, Lauri Carlsson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund & Anssi Yli-Jyrä (red.): *Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*. CSLI Studies in Computational Linguistics ONLINE: CSLI Publications, 71–83.

Koskenniemi, Kimmo. 1983. *Two-level morphology: a general computational model for word-form recognition and production.* Publication No. 11. Helsinki: University of Helsinki Department of General Linguistics.

Kučera, Henry & W. Nelson Francis. 1967. *Computational Analysis of Present-day American English*. Providence: Brown Unviersity press.

Levine, R.D. & W.D. Meurers 2006: Head-Driven Phrase Structure Grammar. Encyclopedia of Languaged & Linguistics, 237–252. http://dx.doi.org/10.1016/B0-08-044854-2/02040-X

Moshagen, Sjur, Pekka Sammallahti & Trond Trosterud. 2004. Twol at work. I Arppe Antti, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund & Anssi Yli-Jyrä (red.): *Inquiries into Words, Constraints and Contexts*. Stanford, California: CSLI, 94–105.

Moshagen, Sjur, Thomas Omma & Tomi Pieski. 2008. *Goallosteapmi Divvunreaidduin.* Tromsø: Universitetet i Tromsø. – http://giellatekno.uit.no/background/Goallosteapmi_Divvun.pdf

Moshagen, Sjur N., Tommi A. Pirinen & Trond Trosterud. 2013. Building an open-source development infrastructure for language technology projects. I *Proceedings of the 19th Nordic Conference of Computational Linguistics* NODALIDA 2013. (NEALT Proceedings Series 16:343–352).

Nickel, Klaus Peter & Pekka Sammallahti. 2011. *Nordsamisk grammatikk* Karasjok: Davvi Girji.

Sammallahti, Pekka. 1989. *Sámi-suoma sátnegirji = Saamelais-suomalainen sanakirja*. Ohcejohka : Jorgaleaddji

Tapainen, Pasi. 1996. *The Constraint Grammar Parser CG-2.* Publications of the Department of General Linguistics, 27. Helsinki: University of Helsinki.

Trosterud, Trond. 2002. Morfologiija rolla sámi giellatenologiijas. *Sámi dieđalaš áigečála* 1/2002, 90–105.

Trosterud, Trond & Linda Wiechetek. 2007. Disambiguering av homonymi i nord- og lulesamisk. Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007. *Suomalais-Ugrilaisen Seuran Toimituksia* 253. Helsinki: Suomalais-Ugrilainen Seura, 401–421.

Trosterud, Trond & Heli Uibo. 2005. Consonant gradation in Estonian and Sámi: two-level solution. I Antti Arppe, Lauri Carlsson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund & Anssi Yli-Jyrä (red.): *Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*. CSLI Studies in Computational Linguistics ONLINE: CSLI Publications, 136–150.

Visl-Group. 2008. *Constraint grammar.* Odense: University of Southern Denmark. http://beta.visl.sdu.dk/constraint grammar.html.

Summary
The article presents a program for analysing North Saami running text, consisting of a morphological transducer and a constraint grammar for disambiguation and syntactic analysis. The article shows what the program can tell us about North Saami grammar when analysing a North Saami text corpus consisting of a total of 25 million words. We look at the productivity of some central word formation processes for nouns and verbs. Compounding is far more productive than derivation, and of the common compounding types the one with first part in the nominative singular is by far the most productive. We also use the transducer to measure the different grammatical categories of nouns and finite verbs, and examine how genre dependent these categories are. Despite its size the corpus is not well balanced for research on finite inflectional categories, and such research must take this imbalance into account. For morphological case the situation is different, here the distribution is independent of genre. We see this as reflecting the nature of the different categories: Person-number and tense reflects the participants and their relation in time, and is thus dependent upon the text, whereas case expresses the position of the arguments within the sentence, independent of genre.

Lene Antonsen og *Trond Trosterud*
Institutt for språk og kultur
UiT Norges arktiske universitet
9037 Tromsø
lene.antonsen@uit.no, trond.trosterud@uit.no