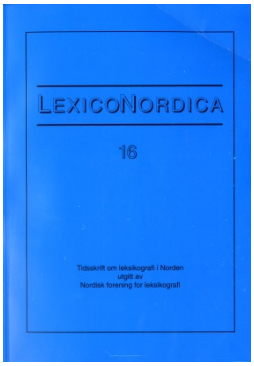
LexicoNordica

Titel: Ei intelligent elektronisk ordbok for samisk

Forfatter: Lene Antonsen, Ciprian-Virgil Gerstenberger, Sjur Nørstebø Moshagen & Trond Trosterud

Kilde: LexicoNordica 16, 2009, s. 271-283

URL: http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Søgbarhed
Artiklerne i de ældre LexicoNordica (1-16) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Lene Antonsen, Ciprian-Virgil Gerstenberger, Sjur Nørstebø Moshagen & Trond Trosterud

Ei intelligent elektronisk ordbok for samisk

The article presents *Vuosttáš Digisánit* (VD), an electronic dictionary from North Sámi to Norwegian. Its novelty lies in the way we have utilized existing resources (a basic dictionary and a morphological analyser/generator) in order to create a reception dictionary for language learners for a morphologically rich language. With only 7,9 % of the word forms in Sámi running text being identical to the lemma form, an approach along the lines sketched here is a prerequisite for a text-integrated e-dictionary. Being a learner dictionary, VD also gives key paradigms for each lemma. This paradigm is generated when building the dictionary, using our language technology tools. We have also built an infrastructure that can be reused for other languages and dictionaries. Our approach shows how it is possible to build text-integrated electronic dictionaries for morphologically complex languages with limited means. The dictionary is available free of charge at:
http://giellatekno.uit.no/words/dicts/index.eng.html.

1. Innleiing

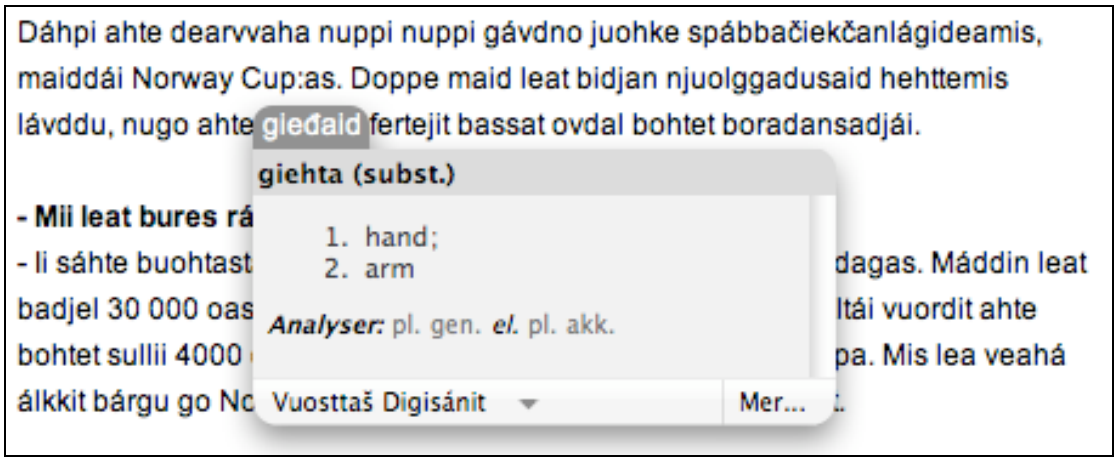
Denne artikkelen presenterer den første elektroniske ordboka frå nordsamisk til norsk, *Vuosttáš Digisánit*, 'Dei første digiorda' (VD). Ordboka inneheld ein morfologisk analyse av det samiske ordforrådet med ei basisordbok for samisk, og resultatet blir ei resepsjonsordbok til bruk ved lesing på skjerm. For språk med rik morfologi er dette i praksis den einaste måten å få ei fungerande elektronisk resepsjonsordbok på, og resultata våre er dermed relevante også ut over samisk.

VD er tilgjengelig på internett for gratis nedlasting på adressa: *http://giellatekno.uit.no/words/dicts/index.nno.html*.

2. Ordboka «Vuosttáš Digisánit»

2.1. Ein kort presentasjon av korleis ordboka fungerer[1]

VD er ei ordbok som kan brukast ved lesing av tekst. I ein nettartikkel frå avisa Ávvir slår vi t.d. opp på ordforma *gieđaid*[2], ved å føre musepeikaren over ordet og trykke på ein tastekommando. VD gjev grunnform, omsetjing, og moglege morfologiske analyser, som vist i Figur 1.



FIGUR 1. Oppslag i teksten, på ordforma *gieđaid*.

Viss det ikkje er nok, har brukaren høve til å trykke på knappen **Mer...** . Då opnar maskina ordboksprogrammet og gjev den fulle artikkelen for *giehta*. I tillegg til den same informasjonen får vi eit miniparadigme (nok til at brukaren skal kunne konstruere det fulle bøyingsparadigmet for *giehta* sjølv), og eventuelle eksempelsetningar. Jf. Figur 2:

---

1 Skildringa er basert på versjonen for MacOS X, men tilsvarande funksjonalitet finst for Windows og Linux, sjå lenger ned i artikkelen.

2 Teksten i dømet er tilgjengeleg på http://www.avvir.no/index.php?news=619.

FIGUR 2. Full ordboksartikkel for *giehta*

2.2. Bakgrunn

Utgangspunktet for arbeidet med VD var at vi hadde tilgang til leksikografiske ressursar, ein morfologisk generator og eit brukargrensesnitt for ordbøker. Ein opplagt idé var dermed å kombinere desse tre ressursane til ei elektronisk ordbok sånn som vi meinte ho burde sjå ut, nemleg som eit verkty der brukaren kan slå opp ord direkte i løpande tekst. Ut over tilgangen til ressursar ser vi også at det veks opp ein ny generasjon lesarar som ikkje er vane med å slå opp i trykte ordbøker men som er vane med lett og direkte tilgang til informasjon. Sjølv for vande ordboksbrukarar er ei morfologisk e-ordbok hurtigare å slå opp i enn ei vanleg ordbok, og for nye brukargrupper kan ei ordbok som VD vere den einaste dei er viljuge til å bruke.

Dei leksikografiske ressursane var dels Nils Jernslettens *Álgosátnegirji* [3] og dels dei pedagogiske leksikona frå *oahpa.uit.no* (Antonsen et al. 2009). Dei morfologiske ressursane var den samiske morfologiske analysatoren og generatoren utvikla som eit samarbeidsprosjekt mellom *giellatekno.uit.no* og *divvun.no*. Denne analysatoren / generatoren er ein endeleg tilstandsautomat, som inneheld eit i prinsippet komplett sett av nordsamiske leksem[4], kombinert med bøyings- og avleiingsmorfologi, og ein eigen tilstandsautomat for morfofonologiske prosessar[5]. Brukargrensesnittet vi tok utgangspunkt i, var programmet *Ordliste* på MacOS X. Frå og med MacOS 10.5 er det mogleg å leggje til eigne ordbokssressursar. Dette inneber at nye leksikalske ressursar automatisk fungerer som frittståande ordbok, og i ein del innebygde ordbokstenester, m.a. høve til å peike på eit ord, og få opp omsetjingar eller definisjonar direkte på skjermen, uavhengig av kva for program du er i[6].

Det leksikografiske grensesnittet i seg sjølv er med andre ord ikkje nytt. Det nye er måten vi har utnytta eit eksisterande rammeverk for eit morfologirikt språk på.

---

[3] Nils Jernslettens ordbok *Álgosátnegirji – samisk-norsk ordbok* vart laga i 1983, med ny utgåve i 1988 utvida med fleire lemma. I presentasjonen av ordboka står det at ho er laga for bruk i undervisninga av samisk rettskriving, men at ho også kan bli brukt i undervisninga i samisk som framandspråk (Jernsletten 1988:7–8). Boka er framleis veldig mykje brukt i begynnaropplæringa, og ho er trykt i fleire opplag. Vi takkar forfattaren for å gjere ordboka tilgjengeleg for oss.

[4] Den endelege tilstandsautomaten inneheld alle leksem vi har funne i korpuset vårt og tilgjengelege ordbøker. Det vil likevel sjølvsagt vere både nyord og andre ord vi ikkje har fått med oss, og korpuset omfattar ikkje all samisk tekst. Men i prinsippet dekkjer tilstandsautomaten alt vi kjenner til.

[5] Generatoren er tilgjengeleg på http://gielletekno.uit.no, både via eit nettgrensesnitt og som nedlastbar open kjeldekode. Jf. Moshagen, Sammallahti & Trosterud (2004) for ein presentasjon.

[6] Det er berre høve til å slå opp i såkalla Cocoa-program, dvs. programmert i eitt av dei tre ulike programrammeverka i Mac OS X. Dette inneber at ordboksprogrammet kan gje oppslag direkte i nettlesaren Safari, men ikkje i Firefox, i redigeringsprogram som Pages og SubEthaEdit, men ikkje i Microsoft Word eller OpenOffice. Cocoa er standardrammeverket for Mac-program, så dei fleste programma for MacOS X er skrivne i det. Men det finst altså ein del svært synlege unnatak.

2.3. Teknisk oppbygging

Teknisk sett finst det fleire måtar å byggje ei elektronisk ordbok på. I den enklaste forma er oppslagsord og søkjeord identiske, slik at ein berre kan søkje på oppslagsforma av eit ord. Eit steg vidare er å leggje til alle bøyingsformer av eit ord som søkjeord for oppslagsordet. Dette tyder at når brukaren slår opp ei ordform, blir ordforma automatisk knytt til lemma, og den tilhøyrande ordboksartikkelen blir vist. Den fyrste versjonen av VD var slik.

Samisk har ein rik morfologi med omfattande morfofonologiske prosessar, som fører til delvis store skilnader mellom grunnform og ein del ordformer. Det å gå frå *vuimmiidanguin* (komitativ fleirtal, 1p. sg. possessiv) til grunnforma *vuoibmi* 'styrke, makt' er ikkje alltid trivielt for ein som skal lære samisk. For betre å hjelpe slike brukarar, valde vi å lage ei ordbok der både lemma og ordformer har sine eigne artiklar. På det viset kunne vi ta vare på mest mogleg lingvistisk informasjon som vi kunne presentere til brukaren, særleg til språkstudentar. Dette fører til eit format med to slags ordboksartiklar: ein grunnformsartikkel, og ein artikkel for andre ordformer. Begge artikkeltypane viser både ordklasse og alle moglege morfosyntaktiske analyser av den spesifikke ordforma (t.d. *vuoidat* 'smørje' inf. eller indik. pres. 1p. pl.) og ei liste med alle omsetjingane til målspråket. Grunnformsartiklane inneheld i tillegg nøkkelformer for oppslagsordet, og bruksdøme med omsetjingar (sjå meir utførleg skildring lenger ned).

Under kompilering blir kvart lemma drege ut frå ordboksfila og til den nordsamiske ordformgeneratoren.

Resultatet av genereringsprosessen er ei liste over alle ordformer og -analyser for dei opne ordklassene. Lista blir ført saman med grunnordboka og sortert etter ordform og ordklasse. I denne fasen blir det generert peikarar mellom ordform og grunnform.

Det er eitt unnatak: lemmaartiklane for bøygde former av lukka ordklasser kan også innehalde eksempel. Dette er mogleg fordi desse bøygde formene er skrivne inn manuelt, dei er ikkje generert slik som ordformene for dei opne ordklassene.

Paradigmet for eit lemma kan berre bli generert dersom det finst som lemma i den morfologiske generatoren vår, og vi var dermed nøydde til å gjere nokre tilpassingar i generatoren. I VD er det ein del fleirtalssubstantiv som berre fanst i eintal i analysatoren, og for desse måtte vi leggje til fleirtalsforma som lemma. I slike tilfelle har formene ulik tyding, slik som *gávdni* = 'gagn, nytte' vs *gávnnit* = 'sengeklær', *gáfet* =

'kaffepulver' vs. *gáffe* = '(ferdigkokt) kaffe', *gaskabeaivi* = 'midt på dagen' vs. *gaskabeaivvit* = 'middag'.

Nokre lemma er homonyme i grunnforma, men har ulike paradigme. Både for å kunne peike tilbake frå bøygde former til riktig lemma, og for å kunne generere korrekt miniparadigme, må vi skilje lemma frå kvarandre ved hjelp av taggar. Dette må vi gjere i den morfologiske analysatoren.

TABELL 1. Homonyme former med ulik analyse.

Sg. Nom.	Sg. Gen.	Norsk	Fst-analyse
lohkki	lohki	lokk	lohkki+N+Sg+Nom
lohkki	lohkki	lesar	lohkki+N+Actor+Sg+Nom
beassi	beasi	reir	beassi+N+Sg+Nom
beassi	beassi	never	beassi+G3+N+Sg+Nom

I tabellen finst det to døme på taggar som vart lagt til. **Actor** er ein derivasjon av eit verb, der den avleidde forma ikkje har stadieveksling. **G3** viser at stadievekslinga er mellom overlang og lang konsonant utan at dette blir markert i rettskrivinga, slik at nominativ og genitiv eintal blir ortografisk identiske.

VD blir kompilert både for MacOS (for Apple sitt ordboksrammeverk) og i XML Dictionary Exchange Format (http://xdxf.sourceforge.net/). For Windows og Linux bruker vi programmet StarDict. StarDict er open kjeldekode og fritt tilgjengeleg (http://stardict.sourceforge.net/).

Det bør nemnast at grunnen til at vi valde løysinga med å lage artiklar for alle ordformer heilt er styrd av dei grensene som ordboksrammeverka gjev oss. Det ville ikkje vere noko problem å gje same informasjon med berre ein artikkel per lemma dersom det hadde vore mogleg å byggje inn ein morfologisk analysator i ordboka. Då hadde vi både fått all lingvistisk informasjon og lemma på same gong. Dessverre er dette ikkje mogleg i dag, og dermed valde vi i staden løysinga med to typar ordboksartiklar. Det er klart at dette gjev ei elektronisk ordbok der ein svært stor del av informasjonen er repetert, men i og med at byggeprosessen repeterer innhaldet automatisk for oss, betyr det i praksis ikkje noko anna enn at ordboka tek ein god del meir plass på harddisken enn det ho elles ville ha gjort.

2.4. Innhaldet i ordboka

Vuosttaš Digisánit inneheld 5192 lemma, 4304 frå Jernsletten si ordbok, og 888 frå det pedagogiske leksikonet til *oahpa.uit.no* .

Lemmaforrådet kan delast i ord med og utan bøying. Av dei med bøying, var ei lita gruppe så uregelrett at vi laga separate paradigme for dei.

TABELL 2. Lemma og ordformer i VD.

Ord	Lemma	Ordformer
Ord med bøying, generert i kompileringa (V, N, A, Pron, Num)	4728	504 740
Ord med irregulær bøying (visse pronomen, negasjonsverbet)	68	344
Ord utan bøying (Adv, Pr, Po, Pcle, Conj, Subj, Interj)	396	396
Totalt	5192	505 480

Samisk har store bøyingsparadigme for kvart leksem. Til ordboka genererer vi f.eks. 93 ordformer for kvart verb, 124 ordformer for kvart substantiv og 394 ordformer for kvart adjektiv. Nokre ordformer innafor kvart bøyingsparadigme er homonyme.

For å gjere VD nyttig for språkinnlæraren har vi lagt til fulle bøyingsparadigme for negasjonsverbet og nokre pronomentypar der bøyingsmønsteret er uregelrett. Til verb, substantiv, adjektiv, numeral og nokre pronomen genererer vi miniparadigme med nøkkelformer som hjelp for brukarane for å sjå stadieveksling og eventuell diftongforenkling.

Det er også variasjon i val av nøkkelformer innafor kvar enkelt ordklasse. Ikkje alle substantiv blir brukt i fleirtal, og da vil vi heller ikkje generere dei i fleirtal. Ein del verb blir vanlegvis ikkje brukt i første person, og vi vil ikkje generere former som ikkje er vanlege sjølv om det går an å tenkje seg dei brukte i eit eventyr eller liknande. Dette gjer vi for at språkinnlæraren ikkje skal venje seg til å sjå verbet i uvanlege samanhengar, t.d. *ciellat* – 'å bjeffe': *dat ciellá* 'det/den bjeffar', ikkje *mun cielan* 'eg bjeffar'. I tillegg har vi vêr-verb, som blir brukte berre i 3p. sg., f.eks. *arvit* 'å regne' og vi har resiproke verb, som blir brukte berre i total og fleirtal når det ikkje er komplement, f.eks. *deaivvadit* 'å treffast'.

For å gjere nøkkelformene lett tilgjengelege, har vi lagt til kontekst når det er naturleg. Til dømes bruker vi pronomen for å vise person- og talbøying av verb, og tidsadverbial for å vise temporal bøying.

I samisk er det ei eiga attributtform av adjektivet som blir brukt når det står føre substantivet. Det finst ikkje eintydige reglar for å utleie attributtforma frå grunnforma, og derfor blir attributtforma oppgitt i

nesten alle ordbøker. Jernsletten si ordbok markerer attributtforma ved å oppgje henne føre eit substantiv i parentes. Vi har fylgt denne praksisen med å gje ein kontekst. Sidan ikkje alle adjektiv passar til alle substantiv, har vi brukt 129 ulike substantiv, til dømes: *suhkkes vuovdi* (’tett skog’), men *rukses bivttas* (’rødt klesplagg’).

I samisk har numeral både eintals- og fleirtalsformer. Vi har valt å gje kontekst til fleirtalsforma for å vise samsvarsbøyinga mellom numeral og substantiv, f.eks. *guovttit gápmagat – guvttiid gápmagiid* - ’to par sko’, i nominativ og genitiv.

Vi har lagt til eksempelsetningar til ein del av lemmaa. Desse hentar vi frå det nordsamiske korpuset ved Sametinget/Universitetet i Tromsø. Dette er eit tidkrevjande arbeid, og vi har hittil gjort dette berre for ein liten del av leksikonet. For nokre av pronomena har vi sjølve laga døme.

3. Den morfologiske komponenten i e-ordbøker for morfologirike språk

Alternativet til å ha ei ordbok med morfologisk analyse av ordformer er å ha ei ordbok med berre tilgang til lemmanivået. For å simulere ei slik ordbok laga vi ein automat basert på settet av leksem i det nordsamiske korrekturprogrammet *Divvun*, til saman 99071 lemma, dvs ein grunnformsautomat. Vi laga tilsvarande automatar for finsk (med Suomen Kielen Perussanakirja, Haarala m.fl. 2001) og bokmål (med Bokmålsordboka, Landrø og Wangensteen 1986). Desse automatane let vi analysere eit testmateriale som inneheldt berre korrekt skrivne ord utan eigennamn, jf. tabell 3.

TABELL 3. Dekningsgrad for lemmabasert automat (samisk, finsk, norsk)

	Samisk	Finsk	Norsk
Ordformtyper i testmaterialet	252 461	45 144	64 994
Lemma i automaten	99 071	94 111	38 983
Dekningsgrad for automaten	7,9	10,0	30,5

For samisk kjende lemmaautomaten att berre 7,9 %. Det vil seie at 92 % av ordformtypane i den samiske teksten er forskjellig frå oppslagsforma. For finsk var situasjonen omtrent den same, ein av ti ordformer i løpande tekst er identisk med grunnforma. For norsk, med langt mindre morfologisk markering, er nesten ein tredjedel av ordformene identiske med moglege grunnformer. Merk at det reelle talet for norsk er lågare,

t.d. vil presensformer av verb (*skriver, leser*) feilaktig bli identifisert som substantiv.

Mange av dei bøygde formene i samisk er transparente, mens andre vil by på større utfordringar for ein brukar som er avhengig av ei tradisjonell ordbok. Nordsamisk har to morfofonologiske prosessar som kan gjere det vanskeleg for ein språkinnlærar å vite kva leksem ordforma høyrer til. Det er snakk om stadieveksling i konsonantsentrum og diftongforenkling der diftongen i første staving blir påverka av vokalkvaliteten i andre staving. Såleis er ordformene *gullái* og *bođii* 3p. sg. av verba *gullát* og *boahtit*, men det å finne fram til rett infinitiv er langt lettare i det første tilfellet enn i det siste. Med ei elektronisk ordbok som VD er det mogleg å klikke i teksten og få fram rett ordboksartikkel.

For mange språk med meir transparent morfologi vil ikkje den morfologiske analysen vere like avgjerande. Det vil vere mogleg for brukaren å lære seg dei mest vanlege suffiksa og forkorte ordforma ein vil slå opp, mens ei tilsvarande ordform på samisk rett og slett vil vere umogleg å finne. Til resepsjon, t.d. når ein slår opp ordformer i ein elektronisk tekst, vil ein òg for andre språk framleis ha stor nytte av ein morfologisk komponent, men då er det ikkje tilstrekkeleg med berre bøyingsformene: ein treng i tillegg produktiv avleiing og samansetjing, slik at ein kan få opp omsetjingar av transparente derivasjonar og samansetjingar. Dette drøfter vi meir i neste avsnitt.

Dei aller fleste språka i verda har ein rik morfologi, og ei klikkbar resepsjonsordbok integrert med elektronisk tekstlesing vil for dei fleste språk vere avhengig av ein morfologisk komponent. Mange språk som t.d. samiske, austersjøfinske og semittiske har også mykje ikkje-segmental morfologi, som gjer det vanskeleg å finne fram til grunnforma. For språk med prefiksdominert morfologi (som t.d. bantuspråka) vil det vere vanskeleg å bruke alfabetisk organiserte ordbøker. For desse språka vil e-ordbøker med ein morfologisk komponent vere avgjerande for at brukaren skal kunne finne oppslagsordet i det heile.

4. Vurdering og vidareutvikling av *Vuosttáš Digisánit*

Arbeidet med VD og med samiske e-ordbøker generelt, er framleis på eit tidleg stadium. Vi drøftar her ein del område som treng vidareutvikling. Ein opplagt veg til ei betre ordbok vil vere å utvide lemmaordforrådet. I og med at det ikkje er spesielt for temaet for denne artikkelen (morfologi i e-ordbøker) vil vi ikkje gå inn på dette her, men

arbeid med ordforrådet vil vere ei prioritert oppgåve i det vidare arbeidet med VD.

4.1. Samansette ord

Morfologien i VD er avgrensa til bøyingsmorfologi. Dette har vi gjort av praktiske grunnar: dei rammeverka vi har tilgjengelege gjer det ikkje mogleg for oss å leggje ved analysatoren vår i den ferdige ordboks-pakka. Dermed kan vi ikkje analysere samansette ord, og då er vi heller ikkje i stand til å gje brukarane hjelp for slike ord.

Samtidig er det opplagt at ein slik mekanisme vil vere til stor hjelp for brukarane. Av dei 246 vanlegaste lemmaa som VD ikkje kjente att, var 28,5 % samansette ord. Dette er ikkje uventa: i eit korpus på 1,1 million ord utgjer samansette ord 26,7 % av substantiva, 7,0 % av det totale tekstmaterialet.

Samisk har, som mange andre språk, ofte ei eiga form på substantiva når dei blir brukte i samansetjingar. For ein del substantiv blir den trykk-lette, finale vokalen til forleddet i samansetjingar svekka (*i* > *e*, *u* > *o*). Dermed er det ikkje nok å dele ordet mekanisk i to. Eit ord som *bargojoavku* 'arbeidsgruppe' er samansett av substantiva *bargu* og *joavku* – noko isolert substantiv **bargo* finst ikkje.

Eit motargument mot å tilby dynamisk analyse av samansette ord er at ein risikerer overgenerering, slik at ein presenterer heilt irrelevante ordboksartiklar for eit ord som analysatoren feilaktig trur er samansett. Dette er særleg eit problem for språk som norsk, svensk og dansk, som har mange einstava substantiv på 2–3 bokstavar. For samisk vil dette problemet vere mindre, fordi det er langt færre slike ord: substantiv på nordsamisk har minst to stavingar, og den minste ordlengda ein då får, er VCV, dvs tre bokstavar. Det er få slike ord, og om ein kombinerer Karlssons lov for samansetjing (Karlsson et al. 1995) med ein vekta analysator som alltid gjev den enklaste analysen og samtidig berre inneheld ordtilfanget i ordboka, vil ein sannsynlegvis langt på veg unngå problemet med overgenerering.

Så lenge rammeverka som er tilgjengelege for oss, ikkje gjer det mog-leg å byggje inn ein analysator slik det er skildra over, blir dette for-betringspotensialet heilt teoretisk. Vi har difor ikkje undersøkt nærmare problemet med overgenerering for dynamisk samansette ord.

4.2. Deriverte ord

VD inneheld heller ikkje avleiingsmorfologi, men biletet er litt meir samansett når det gjeld avleiingar.

Ein gjennomgang av dei 1000 mest frekvente ordformene i korpusmaterialet viste at 136 av dei ikkje vart kjende att av VD. Av desse utgjorde passivformene 8,8 %. I løpande tekst er passiv ikkje så frekvent: Vi undersøkte eit nordsamisk korpus på 1,1 millionar ord, og fann at passive former utgjer 1,6 %, og andre dynamisk deriverte former 1,1 %. Viss vi ser på ordformstypar, blir det same korpuset redusert til ca. 150.000 ord, og dei tilsvarande tala er 2,3 % passive former og 4,3 % andre dynamisk deriverte former.

Passiv er i samisk ein avleiingskategori, og ikkje ein bøyingskategori, og passivformer blir dermed ikkje generert til ordboka. Bøyingsparadigmet for passive verb er like stort som for aktive, slik at det å leggje verb med passivderivasjon til dei aktive grunnverba vil fordoble talet på verbformer.

Blant dei 1000 mest frekvente ordformene er det fem aktioformer. Seks av dei 1000 ordformene har påfølgjande enklitisk partikkel, f.eks. *leago* i staden for *lea go* 'være (indikativ presens 3p. sg.) + spørjepartikkel' og har derfor ikkje blitt kjent att av ordboka. I samisk er det mange ulike enklitiske partiklar, og dei kan leggjast til nesten alle ordformer. Det å leggje til t.d. dei ti vanlegaste enklitiske partiklane vil føre til ei tidobling av ordboka. I Noreg er ikkje problemet med enklitiske partiklar stort, fordi dei vanlegvis blir skrivne som eige ord etter vertsordet. I det same nordsamiske korpuset var talet på ordformer med enklitiske partiklar 0,6 %. Ei frekvensordliste laga med basis i tekstar skrivne i Finland vil nok gje eit anna bilete, da det er vanlegare å samskrive dei der enn i Noreg[7]. Ei undersøking av eit mindre samisk korpus frå Finland (6700 ord) indikerer at frekvensen for slike ordformer er dobbelt så høgt (1,2 %), men eit breiare sjangerutval kan gje høgare tal.

Konklusjonen er at ei realistisk forbetring av VD vil vere å leggje til derivasjonane for passivformer. Dette ville meir enn fordoble talet på verbformer. For å unngå at ordboksfila blir for stor, kunne ein i staden fjerne former med possessive suffiks for substantiv og adjektiv, fordi dei er lite i bruk i skriftleg språk, og dermed er overrepresenterte i ordboka

---

[7] Finsk har same type enklitiske partiklar som samisk, og dei blir *alltid* samanskrivne med vertsordet.

samanlikna med nytteverdien for brukaren. I lista over dei 1000 mest frekvente ordformene er slike former t.d. ikkje representert i det heile.

4.3. Disambiguering i kontekst

Som vist i dømet for *giehta* ovafor, inneheld ikkje ordboka disambiguering av ord i kontekst. Ordboka tilbyr både genitiv og akkusativ som moglege analyser (desse formene er alltid homonyme for substantiv), sjølv om ein syntaktisk analyse av setninga vil vise at korrekt analyse er akkusativ. Innafor giellatekno-prosjektet har vi analyseverkty til å løyse opp slik homonymi (jf. Antonsen et al. 2009), men ordboksgrensesnitta vi har tilgjengelege no gjer det ikkje mogleg å integrere syntaktisk analyse med ordboka.

Dette er likevel ikkje noko stort problem for samisk. Rett nok inneheld samisk løpande tekst mykje homonymi, men berre i marginale tilfelle høyrer dei homonyme formene til ulike leksem. Som regel er dei, som i tilfellet med *giehta*, ulike bøyingsformer av same leksem. For språk der homonymi på tvers av ordklasser er meir vanleg (som t.d. for norsk og engelsk), vil ein slik funksjon innebere eit stort leksikografisk framsteg. Med eit slikt system vil ordboka lese heile setninga, disambiguere for ordklasse, og gje berre den relevante omsetjinga.

5. Konklusjon

Denne artikkelen har vist at det er mogleg å lage ei elektronisk resepsjonsordbok med ein direkte oppslagsfunksjon ved å kombinere ei strukturert ordbok med ein morfologisk generator og eit eksisterande grensesnitt for elektroniske ordbøker. Som eksempel har vi presentert arbeidet med *Vuosttáš Digisánit*, ei nordsamisk-norsk nybyrjarordbok.

Arbeidet med ordboka vart styrt av dei leksikografiske og grammatiske ressursane vi hadde, målgruppa (nybyrjarar) kom som eit resultat av dette. Når det er sagt, er prosenten av språklærande (samanlikna med heile språksamfunnet) relativt høgt for samisk, så eit slikt fokus er ikkje urimeleg.

Ordbøker er viktige for små språksamfunn som for store, og elektroniske ordbøker er viktige hjelpemiddel mellom anna i språklæringssamanheng. Det er eit stort arbeid å skrive ordbøker, og ofte er det mangel på kvalifisert arbeidskraft. Eitt av måla med *Vuosttáš Digisánit* var å byggje ein infrastruktur for å publisere ordbøker i elektronisk

form, enkelt og utan ekstra kostnader. Vi håpar arbeidet vårt kan vere til inspirasjon for tilsvarande prosjekt.

6. Referansar

6.1. Ordbøker

Jernsletten, Nils 1988: *Álgosátnegirji. Samisk-norsk ordbok*. Universitetsforlaget.

Haarala, Risto m.fl. 2001: *Suomen kielen perussanakirja*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 55. Helsinki.

Landrø, Marit Ingebjørg og Boye Wangensteen 1986: *Bokmålsordboka. Definisjons- og rettskrivningsordbok*. Oslo/Bergen/Stavanger Trondheim.

6.2. Annan litteratur

Antonsen, Lene, Saara Huhumarniemi and Trond Trosterud 2009: Interactive pedagogical programs based on constraint grammar. *Proceedings of the 17th Nordic Conference of Computational Linguistics. Proceedings Series 4.*

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä and Arto Anttila 1995: *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing 4. Berlin: Mouton de Gruyter.

Moshagen, Sjur, Pekka Sammallahti and Trond Trosterud 2004: *Twol at work. Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*, 94–105. CSLI Studies in Computational Linguistics. http://cslipublications.stanford.edu/koskenniemi-festschrift/10-moshagen-sammallahti-trosterud.pdf

Lene Antonsen, Ciprian-Virgil Gerstenberger, Trond Trosterud
Giellatekno
Fakultet for humaniora, samfunnsvitskap og lærarutdanning
Universitetet i Tromsø
lene.antonsen@uit.no, ciprian.gerstenberger@uit.no, trond.trosterud@uit.no

Sjur Nørstebø Moshagen
Sametinget
Karasjok
sjur.moshagen@samediggi.no