# Ulike veger fra nordsamisk til lulesamisk

*Trond Trosterud, Lene Antonsen, Inga Lill Sigga Mikkelsen & Anders Lorentsen*

The present article evaluates different approaches to the construction of a North Saami-Lule Saami dictionary. The pivot method (via Norwegian) gives rise both to a low coverage and to a massive overgeneration and requires other methods in order to distinguish between good and bad candidate pairs. We show that it is possible to distinguish reliably between good and bad pairs without additional semantic resources, only by using grammatical analysers, a transliteration component and a grammatical machine translation (MT) system. Similar resources are available also for other Saami languages, and the results shown here are likely to be portable also to other similar language pairs.

## 1. Innledning

Ordbøker blir i beste fall skrevet i den takta det er bruk for dem. For minoritetsspråk var den tradisjonelle ordboka sosialantropologens store oversikt over det tradisjonelle ordforrådet, med ordforklaringer på et internasjonalt språk. Etter hvert som minoritetsspråk har blitt tatt i bruk i samfunnsmessige sammenhenger, har vi fått ordbøker fra majoritetsspråk til minoritetsspråk. Økt bruk innebærer også kontakt ulike minoritetsspråksamfunn imellom. I mange tilfeller er disse språka i slekt med hverandre, men de eneste leksikalske ressursene de har til rådighet, er ordbøker til og fra majoritetsspråket.

Vi vil her se på og evaluere ulike strategier for å bygge ei ordbok direkte mellom de to minoritetsspråka nordsamisk og lulesamisk. For begge språka finnes det ordbøker via norsk, men det finnes ikke ordbøker språka imellom. Vi setter opp følgende problem-

stillinger: Hvor store intrasamiske ordbøker kan man lage ved å gå via et majoritetsspråk som pivotspråk (her: nord- til lulesamisk via norsk)? Kan maskinoversetting være nyttig? Er dette metoder man kan bruke også for andre språkpar?

Vi ser først nærmere på bakgrunnen for dette arbeidet og på hva som er ordboksituasjonen for samiske språk, og hvilke typologiske forskjeller det er de samiske språka imellom, og mellom dem og majoritetsspråka. Vi presenterer deretter materialet vi har brukt, og metodene vi har brukt for å behandle det. Deretter kommer en analyse av resultatet og til slutt en konklusjon.

## 2. Bakgrunn

### 2.1. Hvorfor ordbøker mellom samiske språk?

Hvis man ønsker kommunikasjon mellom talere av ulike samiske språk uten at de må skifte over til majoritetsspråket, må det finnes ordbøker mellom de samiske språka. Slike ordbøker vil også gjøre det enklere for samisktalende å lære seg andre samiske språk. Det er en omveg å måtte gå via majoritetsspråket. Norsk og svensk er typologisk svært forskjellige fra samisk, og er dermed ofte en omveg mellom ord for begreper i samisk språk og kultur.

Alle samiske ordbøker har hittil vært mellom et samisk språk og et majoritetsspråk. Det mangler enspråklige samiske ordbøker, og det mangler ordbøker mellom samiske språk. For ordbøker mellom samiske språk finnes det ett unntak, Sammallahti & Xvorostuxina (1991), ei ordbok med ca. 2900 ordpar[1] mellom nordsamisk og kildinsamisk, publisert rett etter Sovjetunionens fall. I tillegg har terminologisk arbeid for ulike samiske språk gjort det mulig å dra ut 1126 ordpar mellom nord- og lulesamisk fra

1 Med ordpar mener vi lemma og oversettingsekvivalent.

samisk termwiki.[2] Plattforma Apertium inneholder flere maskinoversettingsprogram, deriblant også et program fra nordsamisk til lulesamisk med 13 400 ordpar. Denne ordlista er ikke tilgjengelig for brukerne som ordliste, men den kan brukes i arbeidet med å lage ei ordbok.

## 2.2. Tidligere forskning på å lage ordbøker via pivotspråk

Ved å utnytte den transitive egenskapen ved oversetting (A→B og B→C gir A→C) kan man med utgangspunkt i elektronisk tilgjengelige ordbøker generere nye ordbøker basert på et pivotspråk, som for eksempel her: nordsamisk→lulesamisk via nordsamisk→norsk og norsk→lulesamisk, der norsk er pivotspråk. For eksempel vil det nordsamisk-norske ordparet *ahki-alder* og det norsk-lulesamiske *alder-ahke* gi det nordsamisk-lulesamiske paret *ahki-ahke*. For maskinoversettingsprogram som ikke består av modeller strukturert som nevrale nettverk, er transferleksika mellom kilde- og målspråk en sentral del av modellen. I tida før slike nettverk blei tatt i bruk, blei det dermed eksperimentert en god del med ulike måter å lage transferleksika på ved hjelp av pivotmodellen. Også leksikografer har laga ordbøker for nye språkpar ved å gå via eksisterende ordbøker.

På grunn av at ord har flere betydninger, stemmer den transitive relasjonen i pivotoversetting ofte ikke. Hvis vi prøver å lage et nytt språkpar norsk→tysk via engelsk, vil for eksempel norsk *plante* via engelsk *plant* gi tysk *Fabrikanlage* ('fabrikkanlegg') i tillegg til korrekt *Pflanze*, i og med at det engelske ordet *plant* er flertydig. Drøfting av metoder for å skille gode ordpar fra dårlige er sentral i litteraturen om å bygge ordbøker via pivotspråk. Felles for disse metodene er at de tar ulike språklige ressurser til hjelp for å skille mellom de ulike ordpara produsert med pivotmetoden.

normeringsorganet

2 Det terminologiske arbeidet blir primært gjort av *Giellagáldu* og er publisert på Sámi Giellagáldu tearbmawiki.

Nerima & Wehrli (2008) viste at for et ordbokpar spansk-baskisk via engelsk var så mange som 80 % av 100 000 ordpar feil. Metoden de brukte for å validere korrespondansene, var å attestere kandidatene via et parallellkorpus.

For å lage ei engelsk-galisisk ordbok med spansk som pivotspråk brukte Gamallo Otero og Campos (2010) ikke et parallelkorpus, men to korpus henta fra samme domene. Med utgangspunkt i en dependensanalyse av de to korpusa og ei engelsk-galisisk ordliste laga de for hvert verbpar to sett med substantiv som blei brukt som argument for samme verbpar. For par av substantiv produsert med pivotmetoden prioriterte de deretter de substantivpara som blei brukt som argument for samme verbpar. Forfatterne pekte på at validering med parallell argumentstruktur var kompatibelt med tilsvarende validering basert på parallellkorpus, og at metoden deres dermed kunne brukes som erstatning for parallellkorpus.

Aker et al. (2014) drøfta utarbeiding av tospråklige ordbøker mellom 22 offisielle EU-språk basert på setningsparallelliserte korpus. I eksperimentet deres blei setningene ordparallellisert med GIZA++ (Och & Ney 2000), og de resulterende ordpara blei evaluert etter om de hadde et høgt fonologisk samsvar, om de hadde en høg logaritmisk sannsynlighetsrate (Log-likelihood Ratio, LLR), og om de også kunne genereres ved hjelp av et pivotspråk. LLR er en metode som favoriserer ordparkandidater som opptrer ofte sammen. Fonologisk samsvar blei målt ved å utjevne ortografiske forskjeller (som for eksempel engelsk *inflection* ~ fransk *flexion*) og deretter måle Levenshtein-distansen[3] mellom orda i hvert ordpar. Å kombinere alle tre metodene ga best resultat.

For å utarbeide ei ordbok for ett språkpar (islandsk-engelsk) brukte Úlfarsdóttir & Steingrímsson (2022) en kombinasjon av pivotoversetting, parallellkorpus og maskinoversetting. For alle

3 Levenshtein-distanse forteller hvor mange forandringer man må gjøre for å få et ord til å bli et anna. Levenshtein-distansen mellom *mat* og *mot* er 1, og mellom *bil* og *båt* er den 2 (jf. Levenshtein 1966).

tre metodene blei det brukt flere ulike ressurser, og de ordparkandidatene som fikk best resultat samla, blei valgt ut, likevel slik at ordbokredaktørene måtte skjære ned i materialet. Av de ulike metodene var pivotmetoden den beste. Over 85 % av ordpara som blei valgt, var resultat av pivotoversetting. Aller best resultat fikk de ved å kombinere pivotoversetting med andre metoder. Også andre studier viser at kombinasjonen av flere metoder er den beste måten å skille mellom gode og dårlige pivotoversettinger på.

Oppsummert kan vi si at den sentrale metoden for å lage nye ordbøker basert på eksisterende ressurser er pivotoversetting, og at det deretter blir brukt et bredt spekter av tilgjengelige metoder for å skille mellom gode og dårlige kandidatpar.

## 2.3. Sammenlikning av de samiske språka

Samiske språk tilhører den uralske språkfamilien, de har en rik verbmorfologi, 6–9 kasus og mange morfofonologiske prosesser. Åtte av de samiske språka har en egen offisiell ortografi. Store forskjeller i ortografi gjør at også kognater kan være vanskelige å kjenne igjen, og blant kognatene er det også mange falske venner. Nyere lånord gjør avstanden mellom språka større, da slike lånord stort sett kommer fra majoritetsspråket i landet hvor de samiske språka snakkes, dvs. fra norsk, svensk, finsk og russisk. De samiske språkgrensene går fra kysten og inn i landet, og språka forandrer seg i et kontinuum fra de sørligste delene av det samiske språkområdet (Innlandet og Idre) helt ut til Kolahalvøya. Lengre geografisk avstand mellom språka fører til større forskjeller.

Språka nærmest nordsamisk (i Norge, Sverige og Finland) er lulesamisk i sør (i Norge og Sverige) og enaresamisk i øst (i Finland). Nord- og lulesamer kommuniserer til en viss grad med hverandre på hvert sitt samiske språk. Det leksikalske sammenfallet mellom nordsamisk og lulesamisk er ifølge Tillinger (2014: 77, 88) 83,2 %, basert på ei Swadesh-liste på 184 ord. Dette er like

stor avstand som mellom norsk og islandsk (83,7 %). De ortografiske forskjellene er større enn de språklige forskjellene tilsier: De samiske ortografiene kan klassifiseres i en sørlig (latinsk alfabet med digrafer), en nordlig (latinsk alfabet med diakritiske tegn) og en østlig (kyrillisk alfabet) type. Skillet mellom den sørlige og den nordlige typen utgjør samtidig skillet mellom nord- og lulesamisk, noe som gjør at ellers like ord kan bli skrevet svært forskjellig.

Lulesamisk er det nest største samiske språket, mens nordsamisk er det største. Man har ikke eksakte tall på antall samiskspråklige, men Sammallahti (1998:1), anslår at det er 2000–3000 talere av lulesamisk og 17 000 talere av nordsamisk. Sameloven, som trådte i kraft i 1989 og har som formål å sikre at samene i Norge kan utvikle blant anna språket sitt, hadde først bare kommuner med nordsamisk språk med i forvaltningsområdet for samisk språk. Lulesamisk blei først del av forvaltningsområdet i 2006, da tidligere Divtasvuona suohkan / Tysfjord kommune blei innlemma. Forskjellen i antall talere og tidligere offisiell status i Norge påvirker naturlig nok bruken av språka. Lulesamiske språkbrukere som vil delta i samiske samfunnsdebatter og på samiske arenaer, vil trenge ei nordsamisk ordbok.

Bruk av språka og statusen deres gjenspeiler seg tydelig i det samiske tekstkorpuset SIKOR. Her består det nordsamiske korpuset av 38,94 millioner ord, mens det lulesamiske bare er på 1,8 millioner ord. Termer og ord knyttet til tradisjonelle næringer og tradisjonell kultur er i stor grad like i nord- og lulesamisk, mens nordsamisk i mye større grad er blitt tatt i bruk på nye domener. Arbeidet med å utvikle nye termer er kommet lenger for nordsamisk enn for de andre samiske språka (jf. Trosterud & Eskonsipo 2012), og det ville være både lettere og (sett fra et nabospråkperspektiv) ønskelig om terminologiarbeid for de andre samiske språka kunne dra nytte av arbeidet som er gjort for nordsamisk.

## 3. Materiale og metode

Som ved tidligere arbeid med å lage nye tospråklige ordbøker ved hjelp av pivotmetoden tok vi utgangspunkt i to ordbøker, mellom norsk og hvert av de to samiske språka. For å lage ordlister fra nordsamisk til lulesamisk brukte vi også automatisk translitterering fra nord- til lulesamisk, grammatiske analysatorer for nord- og lulesamisk, et nordsamisk-lulesamisk parallellkorpus på 220 000 ord på hvert språk og et maskinoversettingsprogram fra nord- til lulesamisk.

### 3.1. Treff via ordbøker med norsk som pivotspråk

Vi brukte ei foreløpig utgave av Anders Kintels norsk-lulesamiske ordbok (2012). Dette er den eneste ordboka av noen størrelse som finnes mellom norsk og lulesamisk. Ordboka har en norsk-lulesamisk del og en lulesamisk-norsk del, og arbeidet begge veger er gjort manuelt. Lemmaartiklene har bare to felt («lemma» og «oversettelse») og er dermed ikke maskinlesbare. Vi skripta innholdet ved hjelp av programmeringsspråket Python for å få ut flest mulig lemmapar, og vi endte opp med 18 400 ordpar.

For nordsamisk brukte vi ordboka *Neahttadigisánit nordsamisk-norsk* (heretter *Neahttadigisánit*), som inneholder 30 600 lemmaer utenom egennavn (Antonsen, Trosterud & Eskonsipo 2013–2022). Ordboka foreligger i xml-format og er maskinlesbar. Nordsamisk og lulesamisk ligger både typologisk og kulturelt nær hverandre, mens samiske språk og norsk er både typologisk og kulturelt langt fra hverandre, og slik er norsk som pivotspråk en usikker omveg.

### 3.2. Grammatiske analysatorer

I arbeidet brukte vi to grammatiske analysatorer, Giella-sme for

nordsamisk og Giella-smj for lulesamisk.[4] Analysatorene er bygd med et enspråklig leksikon (dvs. ei lemmaliste) og inneholder stier fra hvert lemma til en morfologisk komponent. En egen modul gir en syntaktisk analyse av setninger (se Antonsen & Trosterud 2017). Analysen gir derfor lemma med morfologisk analyse og syntaktisk funksjon i setninga. Giella-sme er best utbygd og har 155 000 lemmaer i leksikonet. Av disse er 126 000 substantiv, adjektiv og verb, og av dem får 77 % også analyse som sammensetning eller ordavledning. De tilsvarende talla for Giella-smj er 47 500 og 28 000, og av den siste gruppa får 50 % analyse som sammensetning eller ordavledning. De samiske språka har i tillegg ei felles ordliste med 35 000 akronymer og egennavn, de fleste på majoritetsspråka, som får kasusendelser i analysatorene. Analysatorene kan også gi analyse til ordavledninger og sammensetninger som ikke er i leksikonet, og disse kaller vi *dynamiske*. Analysatorene brukes i oversettingsprogramma på Apertium-plattforma, og der kan de dermed lage sammensetninger og ordavledninger etter samme mønster som i kildespråket.

Giella-smj er også nyttig i arbeidet for å vurdere om translittererte ord virkelig er lulesamiske ord (se del 3.3). Hvis ordet får analyse, er det et potensielt lulesamisk ord, og hvis ordets grunnform er lista i leksikonet, er det attestert at ordet brukes i lulesamisk. Men fordi analysatoren ikke inneholder hele ordforrådet, er det en viss feilkilde i denne metoden.

## 3.3. Translitterering av nordsamiske ord til lulesamisk

Forskjellen mellom nord- og lulesamisk er større i skrift enn i tale, av to årsaker. For det første har de to rettskrivingene opphav i to ulike ortografiske tradisjoner. Nordsamisk rettskriving går tilbake til Rasmus Rasks (1832) prinsipp om at hver språklyd skulle ha

4 *Giella* betyr 'språk' på nordsamisk, og sme og smj er iso-kode 639-3 for henholdsvis nordsamisk og lulesamisk.

sin bokstav (*č*, *š*, *ž*), mens lulesamisk står i en tradisjon der man ikke gjorde ei slik endring (og har *tj*, *sj*, (*d*)*tj*). For det andre blei det for de nord- og lulesamiske ortografiene (fra henholdsvis 1979 og 1983) valgt ulike strategier. De to viktigste var at final trykklett ikke-lav vokal blir skrevet *i*, *u* på nordsamisk og *e*, *o* på lulesamisk (for eksempel *oassi*/*oasse* 'del' og *veasku*/*væssko* 'veske'), og at ei av de mest sentrale stadievekslingene blir skrevet omvendt, slik at for eksempel ordet for 'draug' blir skrevet *rávga* (genitiv *rávgga*) i nordsamisk og *rávgga* (genitiv *rávga*) i lulesamisk.

For å kunne se bort fra de ortografiske forskjellene lagde vi et translittereringsprogram, i form av en endelig tilstandsautomat. Forskjellen ortografiene imellom er mer kompleks enn denne korte presentasjonen gir inntrykk av. Samsvaret bokstavene imellom er avhengig av både hvilken bøyingsklasse ordet tilhører, og hvor i ordet den aktuelle bokstaven opptrer. Veksling i rotvokalen er i mange tilfeller avhengig av final vokal, som når nordsamisk *ea* tilsvarer lulesamisk *e* foran stammefinal *u* eller *a*, men lulesamisk *æ* foran andre vokaler. Translittereringa blei brukt til å vurdere kvaliteten av ulike ordpar generert via den nordsamisk-norske og den norsk-lulesamiske ordboka. Dette skjedde ved at de genererte ordpara blei sammenlikna med de translitterte formene, se del 4.1.

## 3.4. Bruk av oversettingsprogram

UiT Norges arktiske universitet har utvikla et maskinoversettingsprogram fra nordsamisk til lulesamisk som er regelbasert og finnes på Apertium-plattforma som åpen kildekode (jf. Antonsen et al. 2017, Antonsen & Trosterud 2020, selve programmet: UiT maskinoversetting). Innputt på kildespråket analyseres med Giella-sme, den grammatiske analysatoren for nordsamisk, og deretter oversettes lemmaene via ei tospråklig ordliste som inneholder 13 400 manuelt produserte ordpar i tillegg til akronymer og egennavn som er henta fra analysatorenes leksikon. Den grammatiske ana-

lysen gjør det mulig for systemet å bygge nye ord på målspråket etter samme mønster som innputt. Det vil si at det kan generere ordavledninger og sammensatte ord og kombinasjoner av disse. Programmet inneholder også transferregler som endrer nordsamisk grammatikk til lulesamisk grammatikk. Deretter genereres setninger på målspråket ved hjelp av den lulesamiske analysatoren, Giella-smj.

Oversettingsprogrammet gir to typer ressurser til ei ordbok. Ordlista i programmet inneholder nyttige ordpar, selv om den bærer preg av å være tilpassa konteksten i tekstene som programmet er utvikla på grunnlag av. Programmet gir også forslag til sammensetninger og ordavledninger som ikke er i ordlista.[5]

## 3.5. Bruk av korpus til validering

I flere av forsøka på å lage nye ordbøker ved pivotoversetting (jf. del 2.2) blei det brukt ulike former for parallellkorpus for validering. Når orda *mel*, *elvemel*, *bredd*, *elvebredd*, *sandbanke* og *bank* alle kan være norske forslag til ei pivotoversetting av engelsk *bank*, er det nyttig å bruke parallellkorpus for å se hvilke kontekster de ulike forslaga opptrer i. Et klassisk eksempel er Europarl (Koehn 2005), som allerede for 25 år siden inneholdt setningsparallelliserte korpus på rundt 50 millioner ord per språk for de vesteuropeiske EU-språka. For engelsk-galisisk brukte Gamallo & Campos (2010) et kompatibelt korpus (med kompatible tema heller enn parallell tekst). Også dette korpuset var på rundt 35 millioner ord. Til sammenlikning inneholder korpuset med parallelle nordsamiske og lulesamiske tekster bare 220 000 ord på hvert språk. Disse tekstene er heller ikke oversettinger av hverandre, men oversettinger fra et tredjespråk.

5 Også i arbeidet med å lage en islandsk-engelsk ordbok brukte man ifølge Úlfarsdóttir & Steingrímsson (2022) Apertium-systemet, men bare for å hente ordlista. Som oversettingssystemer (for oversetting av faste flerordsuttrykk) brukte de Google Translate og Microsoft Translator.

Bruk av GIZA++ var i vårt arbeid ikke til mye hjelp, som også andre har erfart (jf. Aker et al. 2014). Det resulterte i 57 300 ordpar, men selv blant de ordpara som var høyest rangert, var tilnærma alt feil. Med et større og mer presist parallellkorpus vil nok resultata bli gode nok til at de kan bidra til å finne gode ordpar, men i vårt tilfelle var de ikke det.

## 3.6. Manuell evaluering

For å evaluere hvor godt de ulike språklige ressursene klarte å skille mellom gode og dårlige ordpar, tok vi 200 tilfeldige ordpar fra lista over ordpar generert med hver metode og evaluerte disse manuelt. Evalueringa blei gjort av en av artikkelforfatterne, som er lulesamiskspråklig lingvist med gode kunnskaper i nordsamisk.

I evalueringa blei ordpara sortert i tre kategorier: 1. ordpar som er nøyaktige, 2. ordpar som må bearbeides før de kan brukes i ei ordbok, og 3. ordpar som er ubrukelige. Underveis i evalueringa av ordpara la vi også til den norske oversettelsen av den nordsamiske delen av ordparet for å lette evalueringsarbeidet. Nord- og lulesamisk har ordpar som er falske venner. For eksempel betyr *moadda* på lulesamisk *mange*, mens tilsvarende *moadde* på nordsamisk betyr *få*. Slike ordpar er svært vanskelige å oppdage om man ikke kjenner til dem, og med støtte fra norsk blir det enklere å identifisere dem. Ved å legge til norsk oversettelse på enten den nord- eller lulesamiske delen av ordparet vil det bli lettere å finne noen til å gjøre evalueringa, da den som vurderer ordpar, kan være sterkest i enten lulesamisk eller nordsamisk. Dette vil være en stor fordel ved en publisering av en nordsamisk-lulesamisk ordbok, da alle ordpar må gjennomgås manuelt.

## 4. Resultater og analyse

Vi tok utgangspunkt i den nordsamiske lemmalista i *Neahttadigisánit* og produserte potensielle lulesamiske kognater med translitterering. Vi prøvde å pare nordsamiske og lulesamiske lemmaer i ordbøkene via norsk som pivotspråk. Nordsamiske lemmaer som ikke lot seg pare med denne metoden, oversatte vi med maskinoversetting. Vi brukte korpus for å finne ut om de maskinoversatte orda var i bruk, og ved å fjerne ord som ikke blei funnet i korpus, høyna vi kvaliteten på listene med potensielle ordpar. Til slutt vurderte vi metoder og kombinasjon av metoder mot hverandre.

### 4.1. Ordpar via ordbøker, med norsk som pivotspråk

Vi brukte ordbøkene presentert i del 3.1 til å generere nordsamisk-lulesamiske ordpar etter pivotmetoden, med norsk som pivotspråk. Som nevnt er det stor kulturell og typologisk avstand mellom norsk og samiske språk, og det gjør at feilraten blir forholdsvis høy når man bruker norsk som pivotspråk mellom nordsamisk og lulesamisk. Det er i stor grad sammenfallende polysemi i nordsamiske og lulesamiske ord. På grunn av denne polysemien har mange av de nordsamiske orda flere oversettinger på norsk. Det norske ordet kan dessuten være flertydig på en annen måte enn det nordsamiske ordet, eller det kan ha et litt anna innhold enn det samiske ordet. Slik øker sannsynligheta for at ordpara er inkorrekte oversettelser mellom de to samiske språka.

Et anna problem oppsto når det innafor et begrepsområde var mange forskjellige ord med spesialisert betydning i nordsamisk og/eller lulesamisk og mange færre ord med en mer generell betydning i norsk. For eksempel var det mange ord for reinsdyr i både nordsamisk og lulesamisk, og disse blei oversatt til norsk med ordet *rein* i tillegg til ei forklaring, som i *hvit rein*. På grunn av

dette blei termene for rein para tilfeldig, og ordboktreff via norsk ga hele 599 ordpar mellom nordsamisk og lulesamisk, hvorav bare en liten andel var korrekte.

Et beslekta problem som også førte til færre ordpartreff, var at mange samiske verb tilsvarer et flerordsuttrykk på norsk. For eksempel blei både nordsamisk *lubmet* og lulesamisk *láddit* oversatt med *å plukke molter*. Slike ordpar kan få treff via norsk hvis ordbøkene bruker samme flerordsuttrykk i oversettinga, og samme ortografi (for eksempel *molter* vs. *multer* eller *multebær*). Også aspektuelle verbavledninger blei oversatt med flerordsuttrykk. For eksempel vil det avleide verbet *váccášit* kunne oversettes til norsk med *å gå seg en tur*, *gå omkring* osv. Sjøl om lulesamisk har et tilsvarende verb, *váttsatjit*, er det ikke sikkert at man finner disse to verba via den norske oversettinga. En del nordsamiske og lulesamiske ord viser til begreper som ikke finnes på norsk. I slike tilfeller inneholdt ordboka bare ei forklaring på norsk, og de samiske orda kunne dermed ikke koples sammen ved hjelp av pivotoversetting.

Av 29 491 nordsamiske lemmaer (substantiv, verb og adjektiv) fra *Neahttadigisánit* fikk 40 % (11 685 lemma) minst en potensiell lulesamisk ekvivalent via ordboktreff med pivotmetoden. Til gjengjeld fikk hvert lemma i gjennomsnitt nesten 3,4 lulesamiske ekvivalenter hver, og her som for tilsvarende arbeid var utfordringa å skille mellom gode og dårlige ordpar. For å gjøre dette sammenlikna vi ordpara med ordpar der den lulesamiske oversettinga blei translitterert med programmet presentert i del 3.3. Til forskjell fra pivotoversetting gir translitterering alltid ett og bare ett lulesamisk ord, og hypotesen vår var at den pivotgenererte kandidaten som lå nærmest den translittererte, ville være den beste.

For de 11 685 orda i den nordsamisk-norske ordboka som fikk minst en lulesamisk ekvivalent via pivotmetoden, translittererte vi lulesamiske kandidater med programmet presentert i del 3.3. Giella-smj gjenkjente 62 % av de translittererte formene. At tallet

ikke var høyere, har flere årsaker. For det første er leksikonet i de to språka ulike. Den lulesamiske kognatforma til nordsamisk *sátni* ('ord') er **sádne*, men denne forma er ikke belagt i lulesamisk og blir dermed ikke gjenkjent av Giella-smj. I lulesamisk brukes ei anna form, *báhko*. For det andre fikk komplekse ord og særlig lånord inkorrekte former i tilfeller der ulike deler av den lulesamiske ordforma var underlagt ulike fonologiske regler. Normeringa av nyere lånordord stemmer heller ikke alltid overens med genereringa fra nordsamisk norm. Translitterering var likevel nyttig for å vurdere kvaliteten til ordpar generert med pivotmetoden, som det går fram av tabell 1.

| | | **Pivotoversettinga** | | |
|---|---|---|---|---|
| 200 tilfeldige ordpar av hver type: | Ordpar totalt | er nøyaktig | må bearbeides | er ubrukelig |
| En oversettelse, identisk med translitterert form | 165 | 100 % | 0 % | 0 % |
| En oversettelse, ikke identisk med translitterert form | 5014 | 77,5 % | 11,5 % | 11 % |
| To oversettelser, ikke identisk med translitterert form | 3787 | 52,7 % | 23,6 % | 24,1 % |
| Flere enn to oversettelser, ingen identisk med translitterert form | 6037 | 27,9 % | 6,0 % | 66,2 % |

Tabell 1: Ordpar oversatt med ordbøker med norsk som pivotspråk og translitterert, evaluering av 200 tilfeldige par i hver klasse (færre når klassen hadde færre enn 200 medlemmer).

I tabell 1 ser vi at translittereringsforma var en viktig indikator på om pivotforma var god. Der det nordsamiske ordet opptrådte i bare ett ordpar som også hadde støtte i den translitterte forma, var alle ordpara gode. For unike ordpar uten slik støtte (dvs. der

ordparet var forskjellig fra det genererte ordparet) gikk andelen ned til 77,5 %, fremdeles et godt resultat. I tilfella med to eller flere oversettelser per nordsamiske ord uten støtte i translitterert form blei resultatet dårligere med flere oversettelser. Der den ene av to oversettelser sammenfalt med den translittererte forma, så vi ingen klare tendenser, og vi så bort fra denne gruppa i evalueringa.

## 4.2. Ordpar via maskinoversetting

I underkant av 18 000 (60 %) nordsamiske lemmaer fra *Neahttadigisánit* fikk ikke lulesamiske ekvivalenter med hjelp av ordboktreff via norsk. Fra denne lista fjerna vi egennavn, akronymer og flerordsuttrykk. Av de resterende lemmaene var 76 % sammensatte ord. En stor del av disse var ord som *alderssammensetning* og *luftfartøy*, som typisk brukes i administrative tekster og viser til begreper som ikke finnes i Kintels ordbok. For en del nordsamiske lemmaer var det brukt norske oversettinger som ikke sammenfalt med det norske lemmaet i Kintels ordbok, sjøl om begrepet var med.

Alle verb, substantiv og adjektiv uten treff blei oversatt med oversettingsprogrammet. Vi la orda inn i ei matrisesetning, for å sikre at programmet analyserte ordet som grunnform og dermed også ga oversettelsen i grunnform (nominativ entall for substantiv og adjektiv, og infinitiv av verb).[6] Nesten 3000 ord ga ingen oversettelse til lulesamisk, fordi noen av delene av ordet ikke fantes i programmet. Et eksempel er *akvakultuvra*, hvor prefikset *akva* ikke finnes som en del som kan gå inn i ei dynamisk sammensetning. 400 ord blei oversatt, men merka med manglende morfologisk generering fordi oversettingsprogrammet ikke kunne generere den typen sammensetninger, for eksempel sammensetninger med superlativ i førsteleddet.

6 Matrisesetning for substantiv og adjektiv: *Dát lea X* (= Dette er X). Matrisesetning for verb: *Son máhttá X* (= Han/hun kan X).

15 254 ord ga oversettelse. 15 % (2270 lemmaer) fantes i den tospråklige ordlista i programmet. Vi ville vurdere oversettingsprogrammets bidrag til å finne nye ordpar, så vi holdt disse utenom evalueringa. Dermed kunne bare dynamiske sammensetninger og ordavledninger få oversetting via denne metoden. Men dette kunne også gi feil resultat, her illustrert med at det norske ordet *busstopp* også kan bli analysert som *buss+topp.* En riktig analyse kan også gi som resultatet et anna ord enn det som er innarbeida i lulesamisk. For å unngå slike dårlige oversettinger ønska vi å få attestert at de lulesamiske orda virkelig finnes i språket. Bare 4 % (490 lemmaer) av de dynamiske oversettingene var i leksikonet i Giella-smj. Ordforrådet i den nordsamiske ordboka inneholder et anna ordforråd enn Kintels ordbok og de andre ordbøkene som blei brukt i arbeidet med å bygge opp leksikonet i Giella-smj (jf. Antonsen & Trosterud 2020:50f.). De resterende oversettingene av dynamiske sammensatte ord blei testa mot det lulesamiske korpuset.

Evalueringa av resultatet blei gjort på et tilfeldig utvalg av ordpar: ei fil med 200 ordpar der det lulesamiske ordet var attestert i korpuset, og ei fil med 200 ordpar uten slik attestasjon. Resultatet presenteres i tabell 2.

| | | **Oversettinga** | | |
|---|---|---|---|---|
| 200 tilfeldige ordpar: | Ordpar totalt | er nøyaktig | må bearbeides | er ubrukelig |
| I korpus | 1710 | 96 % | 3,5 % | 0,5 % |
| Ikke i korpus | 9835 | 87 % | 6 % | 7 % |

Tabell 2: Evalueringa av kvaliteten på ordpar med og uten attestering i det lulesamiske korpuset.

Ordpara i tabell 2 er generert som dynamisk sammensatte ord med oversettingsprogram på Apertium-plattforma. Tabellen viser at det er best resultater for ordpar hvor det lulesamiske lemmaet

blir attestert i korpus, men bare 15 % av ordpara får slik attestasjon. Korpuset er lite, også fordi lulesamisk enda ikke er tatt i bruk som administrativt språk i samme grad som nordsamisk. Mange nordsamiske lemmaer kommer fra oversettinger av offentlig informasjon fra norsk til nordsamisk. Sammensatte ord er ofte transparente og kan settes sammen med samme ledd i alle tre språk. Et eksempel er det nordsamiske ordet *adopterenvirgelohpi* ('adopsjonspermisjon'), som blei oversatt som dynamisk lulesamisk sammensetning til *adoptierimvirggeloahpe*. Kanskje er det enda ikke oversatt noen tekst til lulesamisk om dette temaet. Ei slik oversetting av ordledd gikk likevel galt i noen tilfeller, særlig når det var innarbeida ei anna sammensetning på lulesamisk enn den som er brukt for nordsamisk. I noen tilfeller inneholdt ikke den nordsamiske ordboka det nordsamiske ordet som faktisk er i bruk.

Mesteparten av ordpara kunne vi ikke attestere via korpus, men evalueringa viste at 87 % av de 7138 dynamiske sammensetningene likevel var gode oversettinger fra nordsamisk. 6 % må bearbeides, ved at man i ei ordbok vil gi en kommentar eller restriksjoner på bruken, dvs. at orda i ordparet ikke har helt samme semantikk eller konnotasjoner. Det kan også være ordpar som består av ord som har samme betydning, men hvor man ville ha ønska et synonym som er mer frekvent. 7 % av ordpara blei klassifisert som «ubrukelige». Slike ordpar kan være oppstått ved at et ord ikke finnes i grunnform i ordlista i oversettingsprogrammet, men er homonymt med et anna ord i bøyd form, slik at det blir oversatt som denne ordforma. Sammensetninger kunne få feil analyse ved at førsteleddet i sammensetninger fikk feil kasus.

## 4.3. Vurdering av de ulike metodene

Pivotoversetting via ordbøker til og fra norsk ga bare treff på 39,6 % (11 685) av 29 491 substantiv-, verb- og adjektivlemmaer i

den nordsamisk-norske ordboka. Årsaken kan være liten overlapp mellom de to ordbøkene til og fra norsk, men også språklig variasjon i den norske delen av ordpara. For orda som fikk treff, var det i gjennomsnitt 3,4 treff per lemma (5,3 treff per lemma for de med flere enn ett treff), og der det var mange ordpar, var de fleste para av dårlig kvalitet. Ved å sammenlikne de lulesamiske kandidatene med translitterert former og ved å skille mellom en, to og flere oversettinger per nordsamisk lemma fikk vi sortert ordpara etter kvalitet. Ved å evaluere utdrag på 200 ordpar fra hver gruppe estimerte vi at pivotmetoden kombinert med translitterering ga i underkant av 5200 ordpar med 78 % av god kvalitet, noe som utgjorde en tredel av ordpara som blei funnet. To tredeler av ordpara som blei generert ved pivotoversetting, var altså av dårlig kvalitet. Tre firedeler av lemmaene som ikke fikk treff i ordboka, blei oversatt av oversettingsprogrammet, som gikk direkte fra nordsamisk til lulesamisk, og ikke via norsk. Av ordpara hvor det lulesamiske ordet blei attestert i korpus, var nesten alle gode (96 %), men for resten av ordpara var andelen gode ordpar også høy (87 %).

## 5. Konklusjon

Utgangspunktet for denne artikkelen var å at vi ønska å finne ut om ordboktreff med et majoritetsspråk som pivotspråk og maskinoversetting kan brukes for å lage intrasamiske ordbøker. Vi starta med nesten 29 500 lemmaer (substantiv, adjektiv og verb) fra ei nordsamisk-norsk ordbok, og etter ordboktreff, translitterering, maskinoversetting og treff i korpus, sto vi igjen med ca. 17 200 ordpar hvor evalueringa viste at mellom 78 % og 100 % av ordpara var gode. Av disse 17 200 ordpara kom 30 % fra pivotoversetting og 70 % fra maskinoversetting. Ordpara fra maskinoversetting var alle sammensetninger eller ordavledninger, og alle var mer eller mindre transparente. I lista med pivotoversettinger, var bare 11 %

sammensatte ord, og denne metoden ga i større grad ordpar fra grunnordforrådet. Disse 17 200 ordpara vil kunne fungere som et godt utgangspunkt for arbeidet med ei ordbok fra nordsamisk til lulesamisk. Disse kommer i tillegg til den tospråklige ordlista med 13 400 ordpar i maskinoversettingsprogrammet som fantes før dette arbeidet.

Sammenlikna med andre tilsvarende arbeid i litteraturen, har arbeidet vårt hatt få ressurser til rådighet. Vi har hatt ordbøker til og fra samme pivotspråk, vi har hatt et regelbasert maskinoversettingssystem med et transferleksikon, vi har hatt analyse- og genereringsprogram for de ulike språka, og vi har hatt et program for å skrive nordsamiske ord til lulesamisk ortografi. Derimot har vi ikke hatt tilgang til semantiske ordnett eller ordbøker via andre pivotspråk enn norsk. Vi har hatt et parallellkorpus, men det har vært relativt lite (220 000 ord), og setningene har ikke vært oversettelser av hverandre, men oversettelser fra et tredje språk. Ut over ordbøkene til og fra norsk har vi ikke hatt tilgang til andre semantiske ressurser. Alle ressursene vi hadde tilgjengelig, er til en viss grad også tilgjengelige for språkpara nordsamisk-enaresamisk og nordsamisk-sørsamisk. For nordsamisk-skoltesamisk har vi ikke oversettingsprogram. Denne artikkelen viser at den dynamiske sammensetninga og ordavledninga i oversettingsprogrammet var viktig for resultatet. For å få tilsvarende resultat for andre liknende språkpar bør man vurdere å bygge et oversettingsprogram bestående av et grunnordforråd og et sett av sammensetnings- og avledningsregler. Et anna bruksområde for metoden drøfta i denne artikkelen vil være å lage ordbøker mellom samisk og majoritetsspråk, der hvor det mangler. Et eksempel ville være mellom norsk og skoltesamisk med hjelp av norsk-finsk og finsk-skoltesamiske ordbøker. Felles for disse og tilsvarende prosjekter for andre språk er at det største problemet er mangelen på kvalifisert arbeidskraft. For samiske og andre minoritetsspråk er det stor etterspørsel etter minoritetsspråklige filologer. Ved å bruke metoder som de vi har

drøfta her, vil vi forenkle arbeidet og gjøre det mulig å lage ordbøker for flere språkpar enn det ellers hadde vært menneskelige ressurser til.

Alle de leksikografiske og språkteknologiske ressursene som er brukt her, er basisressurser som alle språk trenger. Muliheta for å få ordbøker mellom minoritetsspråk er for minoritetsspråksamfunn enda en grunn til å lage slike ressurser.

## Litteratur

### Ordbøker, termbaser, korpus og maskinoversetting

Antonsen, Lene, Trond Trosterud & Berit Merete Nystad Eskonsipo (2013–2022): *Neahttadigisánit Davvisámi-dáru-davvisámi sátnegirji.* Tromsø: UiT. <sanit.oahpa.no> (juni 2023).

Apertium. <wiki.apertium.org> (august 2023).

Kintel, Anders (2012): *Julevsáme-dárro báhkogirjje*.

Sámi Giellagáldu tearbmawiki (samisk termwiki). <satni.uit.no/termwiki> (juni 2023).

Sammallahti, Pekka & Anastasija Xvorostuxina (1991): *Unna sámi-sām' sátnegirjjáš = Udc' sám'-sámi soagknegka*. Ohcejohka: Girjegiisá.

SIKOR. *UiT Norges arktiske universitets og det norske Sametingets samiske tekstsamling*, versjon 01.12.2021. <gtweb.uit.no/korp> (juni 2023).

UiT maskinoversetting, på Apertiums plattform. <gtweb.uit.no/mt> (juni 2023).

### Annen litteratur

Aker, Ahmet, Monica Paramita, Märcis Pinnis & Robert Gaizauskus (2014): Bilingual dictionaries for all EU languages. I: *Proceedings of the Ninth International Conference on Language*

*Resources and Evaluation*. *(LREC'08)*. European Language Resources Association (ELRA). Reykjavik, Island, 483–489.

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. I: *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017, Gothenburg, Sweden*. Linköping Electronic Conference Proceedings 131. Linköping: Linköping University Electronic Press, 123–131.

Antonsen, Lene & Trond Trosterud (2017): Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk. I: *Norsk lingvistisk tidsskrift*, Årgang 35:1, 153–185.

Antonsen, Lene & Trond Trosterud (2020): Med et tastetrykk. Bruk av digitale ressurser for samiske språk. I: *Samiske tall forteller* 13. Kommentert samisk statistikk 2020. Kautokeino: Sámi allaskuvla, 43-67.

Gamallo Otero, Pablo & José Ramom Pichel Campos (2010): Automatic Generation of Bilingual Dictionaries Using Intermediary Languages and Comparable Corpora. I: A. Gelbukh (ed.): *Computational Linguistics and Intelligent Text Processing*. CICLing 2010. Lecture Notes in Computer Science, vol 6008. Berlin, Heidelberg: Springer, 473–483. DOI: 10.1007/978-3-642-12116-6_40.

Koehn, Philipp (2005): Europarl: A Parallel Corpus for Statistical Machine Translation. I: *Proceedings of Machine Translation Summit* X. Papers. Phuket, Thailand, 79–86.

Levenshtein, Vladimir I. 1966: Binary codes capable of correcting deletions, insertions, and reversals. I: *Soviet Physics Doklady*, 10 (8): 707–710.

Nerima, Luka & Eric Wehrli (2008): Generating bilingual dictionaries by transitivity. I: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

European Language Resources Association (ELRA), 2584–2587.

Och, Franz Josef & Hermann Ney (2000): A comparison of alignment models for statistical machine translation. I: *Proceedings of the 18th conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1086–1090. DOI: 10.3115/992730.992810.

Rask, Rasmus (1832): *Ræsonneret lappisk sproglære efter den sprogart, som bruges af fjældlapperne i Porsangerfjorden i Finmarken. En omarbejdelse af Prof. Knud Leems Lappiske grammatica*. København: J. H. Schubothes Boghandling.

Sameloven = Lov om Sametinget og andre samiske rettsforhold (sameloven). <lovdata.no/dokument/NL/lov/1987-06-12-56> (juni 2023).

Sammallahti, Pekka (1998): *The Saami languages. An introduction*. Kárášjohka: Davvi girji.

Tillinger, Gábor (2014): *Samiska ord för ord. Att mäta lexikalt avstånd mellan språk*. Studia Uralica Upsaliensia 39.

Trosterud, Trond & Berit Merete Nystad Eskonsipo (2012): A North Sami translator's mailing list seen as a key to minority language lexicography. I: Ruth Vatvedt Fjeld & Julie Mathilde Torjussen (eds.): *Proceedings of the 15th EURALEX International Congress*. Oslo: University of Oslo, 250–256.

Úlfarsdóttir, Þórdís & Steinþór Steingrímsson (2022): Dannelsen af en tosproglig ordbog med hjælp af sprogteknologiske metoder. I: *LexicoNordica* 29, 153–174.

Trond Trosterud
professor
UiT Norges arktiske universitet
NO-9037 Tromsø
trond.trosterud@uit.no

Lene Antonsen
professor
UiT Norges arktiske universitet
NO-9037 Tromsø
lene.antonsen@uit.no

Inga Lill Sigga Mikkelsen
overingeniør
UiT Norges arktiske universitet
NO-9037 Tromsø
inga.l.mikkelsen@uit.no

Anders Lorentsen
overingeniør
UiT Norges arktiske universitet
NO-9037 Tromsø
anders.lorentsen@uit.no