OSLa Hagen, Hjelde, Stjernholm & Vangsnes (red.) *Bauta: Janne Bondi Johannessen in memoriam*, Oslo Studies in Language 11(2), 2020. 489–501. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

HVA ER VIKTIG FOR FORSTÅELSE? OM MASKINOVERSETTING FRA NORDSAMISK

TROND TROSTERUD OG LENE ANTONSEN
UiT Norges arktiske universitet

SAMMENDRAG

Artikkelen presenterer et regelbasert maskinoversettingssystem fra nordsamisk til norsk. Den grammatiske analysen blir gjort med Giellatekno og Divvuns nordsamiske analyseprogram. Vi har skrevet transferkomponenten (transferleksikon og grammatiske regler) innafor rammeverket til det åpne maskinoversettingssystemet Apertium. Artikkelen inneholder ei evaluering av oversatt tekst for to ulike domener. Tekstene skårer bedre på gjengiving av innholdet enn på godt språk. Ved å systematisere feilene i leksikalske, grammatiske og pragmatiske feil, viser vi at leksikalske feil går mest ut over forståelsen. De andre to feiltypene gir dårlig språk, men har liten innvirkning på forståelsen. Feiltypen det er lettest å forbedre, er den leksikalske, noe som er en lovende konklusjon for utviklinga av et maskinovsersettingssystem for tekstforståing.

[1] INNLEDNING

I løpet av det siste halvannet tiåret har statistisk maskinoversetting dominert maskinoversettingsfeltet. Basert først på samsvar på frasenivå og seinere på nevrale nettverk har denne tilnærmingsmåten stort sett overtatt for lingvistisk basert maskinoversetting. Metodologisk sett fins det likevel en liten «gallerlandsby» som holder stand og som baserer maskinoversetting på ENDELIGE TILSTANDSAUTOMATER og FØRINGSGRAMMATIKK, dvs. den teknologien som i si tid gjorde Tekstlaboratoriet i Oslo i stand til å analysere norsk tekst.[1]

Vi vil presentere et program for maskinoversetting fra nordsamisk til norsk bokmål som foregår ved Giellatekno, forskningsgruppa for samisk språkteknologi ved UiT Norges arktiske universitet. Programmet er i daglig bruk, og integrert i så ulike sammenhenger som hjemmesidene til Samisk høgskole og korpuset for samisk talespråk ved Tekstlaboratoriet ved UiO, LIA Sápmi.

[1] Tekstlaboratoriets modell for norsk morfologi er presentert i Johannessen 1990, modellen for syntaks i Johannessen m.fl. 2012. Til sammen utgjør disse to komponentene den såkalte Oslo-Bergen-taggeren.

I og med at teknologien er grammatisk basert, er det også mulig å gi en analyse av hvor godt de ulike grammatiske komponentene i systemet fungerer. Denne artikkelen vil gjøre nettopp det, og vise både hvilke deler av grammatikken som har størst påvirkning på forståelsen av den oversatte teksten, og hvilke deler av systemet man innafor realistiske rammer kan forbedre. Vi kommer til å først presentere den regelbaserte teknologien som ligger til grunn for systemet, før vi viser oppsettet for evaluering og gir en analyse av resultatene. Til slutt kommer en konklusjon.

[2] MASKINOVERSETTING

Maskinoversetting (MT) kan bli brukt på ulike måter: For det første kan vi bruke det for å lese en tekst (for eksempel i form av ei nettside) som er skrevet på et språk vi ikke kan. Kvaliteten på språket i oversettinga er ikke avgjørende – det viktige er at eventuelle feil i oversettinga ikke gir opphav til misforståelser. For det andre kan vi bruke maskinoversetting som hjelp til å publisere teksten vår på et annet språk ved å la maskinoversettingsprogrammet lage en kladd. Oversettingsfeil som gir opphav til misforståelser, er ikke like alvorlige, så lenge de ikke er tidkrevende å rette i korrekturlesinga. Derimot er det viktige at teksten er idiomatisk riktig og så nær opp til publiserbart språk som mulig.

En god oversetting skal gjengi innholdet i originalen og være skrevet på et godt språk. Så langt er alt vel. Problemet oppstår når utviklerne blir tvunget til å velge, eller rettere sagt prioritere mellom disse to målene. Valget er avhengig av måten programmet skal bli brukt på: Hvis målet er å lese for å forstå, bør det å gjengi meningsinnholdet få prioritet, mens maskinoversetting som lager kladd for noe som skal publiseres, bør prioritere idiomatisk riktig språk. I praksis viser det seg at de ulike statistiske metodene har prioritert godt språk, mens regelbaserte system har prioritert nærhet til originalen.

Maskinoversettingsprogram var i mange år regelbasert, men de siste par tiåra har statistiske og deretter nevrale metoder vært de dominerende (Wu m.fl. 2016 gir f.eks. ei framstilling av Googles NMT-system). Disse metodene krever så mye tekst til å trene systemet, at det i praksis er umulig å lage maskinoversettingsprogram for minoritetsspråk.[2] Språkparet vi behandler i denne artikkelen, bruker dermed regelbasert maskinoversetting.

[2.1] *Regelbasert maskinoversetting fra nordsamisk til norsk*

Regelbasert maskinoversetting analyserer kildespråket grammatisk, oversetter ordene i setninga og endrer den grammatiske strukturen fra kilde- til målspråk.

[2] Hassan m.fl. (2018) bruker et parallellkorpus med 100 mill. setningspar mellom kinesisk og engelsk, der vi for nordsamisk-norsk har 108000 tilgjengelige setningspar og for sørsamisk-norsk 6000.

Til slutt genererer systemet målspråksteksten, se eksempel i figur 1.

```
^Prest<n><m><sg><def>$
^ha<vblex><pres>$
^en<det><qnt><m><sg>$
^sønn<n><m><sg><ind>$
```

FIGUR 1: Mellomstadium for oversettinga av setninga *Báhpás lea bárdni.* Her er grunnformene oversatt og taggene for morfologi tilpassa norsk grammatikk. Til slutt blir oversettelsen generert: 'Presten har en sønn'.

Programmet trenger regler for hvert fenomen, også for eventuelle ortografiske eller språklige feil i kildespråket. Når det er flere mulige oversettinger, velger systemet oversettelse av ordet i henhold til regler basert på kontekst (leksikalsk seleksjon). I figur 2 (nedafor) viser vi hvordan det nordsamiske *bárdni* i setninga 'Báhpás lea bárdni' blir oversatt som henholdsvis 'gutt' og 'sønn' utfra kontekst. I en habitivkonstruksjon, som den vi har her, vil 'sønn' bli valgt.

Dette er både styrken og svakheta til regelbaserte system: På den ene sida må alt gjøres eksplisitt for å kunne utføres. På den andre sida har utvikleren også kontroll over resultatet: I og med at systemet er eksplisitt vet vi også hva som kommer ut. Tilpassing av systemet til nye domener er i seg sjøl ikke vanskelig: Den grunnleggende grammatikken er den samme fra domene til domene, og den viktigste tilpassinga innebærer utviding av ordforrådet og nye regler for leksikalsk seleksjon i de tilfellene der det spesifikke fagområdet vil oversette visse ord på andre måter enn det som blir gjort i andre domener.

Ved UiT har vi utviklet oversettingssystem fra nordsamisk til norsk, men også til andre samiske språk (jf. Antonsen m.fl. 2017; Johnson m.fl. 2017).[3] Den morfologiske analysen av kildespråket er gjort med endelige tilstandstransdusere (FST). Ved tvetydighet velger manuelt skrevne regler i føringsgrammatikk (Constraint Grammar) riktig analyse ut fra kontekst, og slike regler legger også funksjonstagger til analysen, se figur 2. (Se Antonsen & Trosterud 2017 for beskrivelse av analysatorene for nordsamisk.)

For sjølve transferkomponenten bruker vi det åpne MT-systemet Apertium (se Forcada m.fl. 2011). Apertium blei opprinnelig utarbeida for oversettelse mellom nære slektspråk, mer spesifikt, språka på den iberiske halvøya. Seinere har flere språkpar kommet til. Alt i alt inneholder Apertium 302 språkpar, også par der språka er svært ulike hverandre, som spansk-baskisk og russisk-engelsk.

Eksempler på Apertium i bruk inkluderer dagsavisene *La Coruña* (spansk-katalansk) og *La Voz de Galicia* (spansk-galisisk) og hjemmesidene til

[3] Programmene er tilgjengelige på http://gtweb.uit.no/mt/. Takk til våre kolleger på Giellatekno og Divvun for godt samarbeid i arbeidet med grammatikkmodellene over ei årrekke, og spesielt til Sjur N. Moshagen som laget oppsettet for å integrere grammatikkmodellene våre i Apertium-systemet.

administrasjonen i Catalunya (katalansk-spansk). På Wikipedia er (3.11.2020) 25468 artikler oversatt med Apertium fra spansk til katalansk, 6202 fra katalansk til spansk, 3314 fra bokmål til nynorsk og 2372 fra nynorsk til bokmål.[4] Det mest brukte Apertium-språkparet er bokmål-nynorsk. Nynorsk Pressekontor bruker Apertium som hjelp for å oversette fra bokmål til nynorsk, og det samme gjør norske skoleelever. Ut over de Apertium-baserte systemene fins det også regelbaserte oversettingsprogrammer mellom norsk bokmål og engelsk via dansk (Bick & Nygaard 2007). Et regelbasert system fra norsk direkte til engelsk blei også i si tid utarbeida (Lønning m.fl. 2004), men det er ikke i bruk.

Arbeid med det nordsamisk-norske maskinoversettingsprogrammet blei igangsatt i 2010 av Kevin Unhammer. Andre utviklere har vært Lene Antonsen, Trond Trosterud og Francis Tyers (se Trosterud & Unhammer 2013 for presentasjon av en tidlig versjon). Målet med programmet er forståelse, dvs. at norskspråklige lesere som ikke kan samisk, skal få tilgang til tekst skrevet på nordsamisk. Dermed blir det mulig for nordsamiskspråklige å skrive på samisk, uten samtidig å miste lesere.

Målgruppe for programmet er altså brukere som ikke forstår nordsamisk. Deres behov er en forståelig tekst som gjengir meningsinnholdet i kildespråket. De vil sjøl kunne vurdere i hvor stor grad oversettelsen er idiomatisk norsk.

Et oversettingsprogram fra norsk til samisk ville ha en annen type målgruppe. Samiske lesere i Norge er tospråklige, og de vil heller lese originalen enn ei oversetting. Programmets brukere ville da være norskspråklige som av forskjellige grunner ville ønske å oversette teksten til et språk de ikke sjøl behersker. Disse brukerne ville ikke kunne vurdere i hvor stor grad oversettelsen er idiomatisk samisk, og slike dårlige oversettinger kunne ukritisk bli publisert på internett. Slik ukorrigert maskinoversatt tekst kunne fort ha utgjort en stor del av den samiske tekstmassen på internett, noe som ville ha vært høyst problematisk for et lite minoritetsspråk.

[2.2] *Kildespråkets lingvistisk analyse*

Figur 2 viser hvordan den maskinelle analysen av kildespråket er utgangpunktet for oversettelsen. Markert med gult er tagger med informasjon om at lokativ kasus (Loc) i den første setninga brukes i en habitivkonstruksjon (<hab>), og i den andre setninga i en stedsbenevnelse (Sem/Plc). Dette utløser nokså forskjellige oversettinger sjøl om setningene ellers er like.

[4] <https://nn.wikipedia.org/wiki/Spesial:Innhaldsomsetjingsstatistikk>

```
"<Báhpas>"
    "báhppa" N Sem/Hum Sg Loc <hab> @ADVL>
"<lea>"
    "leat" V IV Ind Prs Sg3 @+FMAINV
"<bárdni>"
    "bárdni" N Sem/Hum Sg Nom <ext> @<SUBJ
"<.>"
    "." CLB

"<Šiljus>"
    "šillju" N Sem/Plc Sg Loc @ADVL-ine>
"<lea>"
    "leat" V IV Ind Prs Sg3 @+FMAINV
"<bárdni>"
    "bárdni" N Sem/Hum Sg Nom <ext> @<SUBJ
"<.>"
    "." CLB
```

FIGUR 2: To nordsamiske setninger med analyse. På grunnlag av de taggene i analysen som er framhevet med gul farge, blir det genererert to nokså forskjellige norske setninger: 'Presten har en sønn' versus 'På gårdsplassen er det en gutt.'

Samisk har ikke bestemt og ubestemt form, og heller ikke artikkel eller formelt subjekt, så disse må legges til i den norske oversettinga. I den første setninga er adverbialet et ord som refererer til et menneske (markert med taggen Sem/Hum). Her blir konstruksjonen analysert som en eierkonstruksjon (<hab>) og oversatt med 'Presten' som subjekt og verbet 'har'. I den andre setninga er adverbialet et stedsadverbial, og taggen for eksistensialsetning (<ext>) utløser her det formelle subjektet 'det' i den norske oversettinga.

Det er regler basert både på morfologisk, semantisk og leksikalsk informasjon for valg av preposisjoner som skal tilsvare samiske kasus, og i figur 2 velges 'på gårdsplassen' for det samiske ordet i lokativ, andre muligheter ville vært 'i, fra, av, hos'.

Samisk har ikke grammatisk genus, og vi har laget regler som leter etter en antesedent som kan fortelle kjønnet når det brukes personlig pronomen. Hvis det ikke lykkes, velger systemet formen 'hun/han'. Valg av feil preposisjon eller pronomen kan i ulik grad få konsekvenser for hvor forståelig oversettinga blir.

[3] EVALUERING

Vi vil vurdere i hvor stor grad programmet er i stand til å produsere en oversettelse som gjør det mulig å forstå innholdet i den samiske originalteksten, og hva som eventuelt ødelegger forståelsen. For å undersøke dette valgte vi ut to ukjente tekster: En generell avistekst fra den nordsamiske avisa *Ávvir* og en tekst fra Samisk høgskole (SH). Vedlegg 1 viser de originale artiklene.[5] For hver setning stilte vi disse spørsmålene:

[5] Evalueringa blei gjort i september 2018.

(i) Hvor **forståelig** er den oversatte teksten?
(ii) Hvor godt har **innholdet** i setningene blitt bevart fra nordsamisk til norsk?
(iii) Hvilke **språkfeil** er det i oversettingene?
(iv) Hvordan innvirker de språklige feilene på **forståelsen** og **innholdet**?

For å svare på dem hadde vi to evalueringsgrupper, ei gruppe av norskspråklige, som skulle vurdere hvor forståelig og hvor godt språk det var i MT-oversettelsen, og ei gruppe tospråklige, som skulle vurdere hvor godt MT-oversettelsen representerte originalteksten.

[3.1] *Evaluering av målspråket*

De 12 norskspråklige evaluatorene fikk følgende spørsmål for hver setning: Hvor forståelig er den maskinoversatte norske teksten? Det blei gitt fire ulike svaralternativ, som vist i tabell 1.

Vurderingsalternativ	Nyhet	SH-tekst
1. Forståelig setning uten alvorlige grammatiske feil	2	8
2. Tror jeg forstår hele setninga, men håpløs norsk	4	6
3. Forstår deler av setninga	8	0
4. Uforståelig	1	0
Setninger til sammen	15	14

TABELL 1: Norskspråkliges evaluering av målspråket: Hvor forståelig oversettelsen er.

Som vist i tabellen, er det svært stor forskjell mellom nyhetsteksten og SH-teksten. For den sistnevnte var alle setningene forståelige, og for de fleste av dem var språket vurdert som godt. For nyhetsteksten var tvert i mot de fleste setningene delvis eller til og med helt uforståelig, og bare to setninger blei vurdert som ei forståelig setning uten alvorlige grammatiske feil.

Vår vurdering av forskjellen tekstene imellom er at maskinoversetting er avhengig av domenet for oversettinga. Vi har arbeidet mye tekster fra SH-domenet og dette gir utslag i evalueringa. Tema for nyhetstekster er langt bredere, og dermed vanskeligere å dekke.

[3.2] *Evaluering av innholdet til maskinoversettinga*

For å evaluere i hvor stor grad setningens innhold er bevart fra kildespråket til målspråket, trengte vi tospråklige evaluatorer. Her hadde vi 7 evaluatorer. Også

disse fikk fire alternativ, vist i tabell 2.[6]

Vurdering	Nyhet	SH-tekst
1. Innholdet er ganske bra bevart	3	9
2. Deler av innholdet er ikke bevart.	4	4
3. Viktig del av innholdet er ikke bevart	5	1
4. Oversettinga fungerer ikke i det hele tatt	3	0
Setninger til sammen	15	14

TABELL 2: Tospråkliges evaluering av hvor godt meningsinnholdet i originalen er bevart.

[4] DRØFTING

Evalueringa gir oss et utgangspunkt for å si noe om hvilke språkfeil som ødelegger oversettelsen mest. Vi klassifiserte avvikene mellom maskinoversatt tekst og redigert målspråkstekst i leksikalske, grammatiske og pragmatiske feil.

Som leksikalske feil vurderte vi både ord som mangla i systemet, ord som fikk en oversettelse som ikke passa inn i konteksten og feil valg av preposisjoner og pronomener. De grammatiske feilene inkluderte både ordstillings- og bøyingsfeil. Vi skilte mellom grammatiske og pragmatiske feil ved å kalle feil som involverte bruk av bestemthet og determinativ, for pragmatiske feil. I de fleste tilfeller er valget mellom bestemt og ubestemt form av substantiv ikke et spørsmål om grammatikalitet, men om informasjonsstruktur, så vi valgte å se på slike feil som pragmatiske og ikke grammatiske.

Oversettelser representative for hver av de tre typene går fram av oversikten under, hvor eksemplene inneholder leksikalske feil (1), grammatiske feil (2) og pragmatiske feil (3). I parentes er ei mer idiomatisk oversetting.

(1) a. *Gievkkanskábet leat **rámssagan** ja laigan...*
Kjøkkenskapene har ***rámssagan*** (> sprukket) og flaknet...

b. *[...] ja **lasiha** ahte vaikko son lea váldosuodjalusáittardeaddji gieldda bargiid bargodillái*
[...] og **det øker** (> hun legger til) at selv om hun/han er hovedverneombud for arbeidsforholdet til kommunens ansatte

(2) a. *In dieđe **maid** áigot dahkat dainna, ...*
Jeg vet ikke **de** skal **hva** (> hva de skal) gjøre med de, ...

b. ***Buot biebmosuovat...***
Alle matrøykene (> All matrøyken) ...

[6] Disse evaluererne fikk spørsmålene på samisk, alternativa var: Sisdoallu lea seilon oalle bures 2. Oasáš ii leat seilon, 3. Dehálaš oassi ii leat seilon, 4. Ii doaimma ollenge

(3) a. ***Oahpahus*** *lea sámegillii.*
Undervisning (> Undervisningen) er på samisk.
b. *Fágasisdoallu lea čohkkejuvvon* ***fáttáid*** *mielde.*
Faginnholdet er blitt samlet etter **temaene** (> tema)

Evalueringa viste at leksikalske feil hadde stor innvirkning på forståeligheta. De seks setningene som hadde flere enn to leksikalske feil, var også de som skårte dårligst for forståelighet. Tilsvarende fikk de 12 setningene som ikke hadde leksikalske feil, best skår for forståelighet.

Hvis feiltypen var leksikalsk eller grammatisk, førte opphopinga av flere feil samtidig til ei enda dårligere oversetting. I eksempel (4) inneholder oversettinga svært mange feil. I eksempla (4) og (5) er (a) den samiske originalen, (b) ei manuell oversetting av (a), og (c) er MT-versjonen av setninga.

(4) a. *Olles Sámis dárbbašuvvojit oahppan oahpaheaddjit mánáidgárddiin.*
b. Hele Sameland behøver utdannede lærere i barnehagene.
c. Du nå frem i Sameland de behøves utdanne lærere i barnehagene.

Oversettelsen av (4a) inneholder både feiloversettelser og grammatiske feil. Som analyse for *olles* har programmet valgt ei form av verbet *ollit* 'nå fram', i stedet for attributtforma av adjektivet *ollis* 'hel'. Dette har påvirket valg av analyser i resten av setninga slik at *oahppan* 'utdannet' ikke har blitt gjenkjent som perfektum partisipp, men som en nominalisert verbderivasjon som systemet ikke har kunnet overføre til norsk, og derfor blir ordet representert med grunnforma av verbet.

Ei oversetting som inneholder like mange feil, men der feilene er pragmatiske, kan likevel være forståelig. Eksempel (5) viser pragmatiske feil: feil bruk av artikler og bestemthetsformer:

(5) a. *Sámi mánáidgárdeoahpaheaddjeoahpus leat vejolašvuođat oažžut liigestipeanddaid.*
b. På den samiske førskolelærerutdanninga er det muligheter til å få ekstrastipend.
c. På samisk førskolelærerutdanning er det muligheter å få ekstrastipendene.

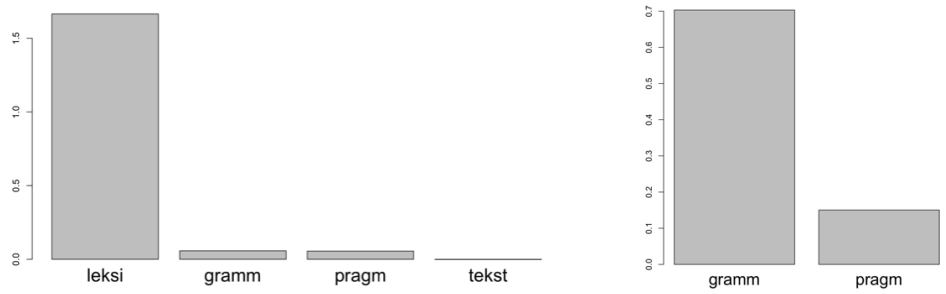
Setninger med bare pragmatiske feil skårte høyt for forståelighet, jf. eksemplene (6) og (7) (her er b-setninga maskinoversatt):

(6) a. *Oahpahus lea sámegillii.*
b. Undervisning er på samisk.

(7) a. *Fágasisdoallu lea čohkkejuvvon fáttáid mielde*
b. Faginnholdet er blitt samlet etter temaene.

For setningene (6) og (7) er problemet bestemthet i maskinoversettinga (b). I (6b) hadde formen *Undervisninga* fungert bedre, og i (7b) *tema* (*fáttáid* er flertallsform). Feilene gir dårligere lesbarhet, men gjør det ikke vanskeligere å forstå innholdet.

Vi utførte også en CART-analyse ('Classification and regression tree') av dataene med informantvurderingene som avhengig variabel og tallet på leksikalske, grammatiske og pragmatiske feil i hver setning som uavhengige variabler. CART gir ei optimal klassifisering av data og måler den relative betydninga av de ulike variablene.[7] Figur 3 viser resultatet av denne prosessen. Figuren til venstre viser at leksikalske feil er langt viktigere for vurderinga enn de andre feiltypene. For å vurdere den relative påvirkninga av grammatiske og pragmatiske feil gjentok vi den samme vurderinga uten den leksikalske faktoren, og effekten av grammatiske feil viste seg å være sterkere enn den av pragmatiske (figuren til høyre).



FIGUR 3: CART-analyse. Til venstre er leksikalske, grammatiske, pragmatiske feil og sjanger uavhengige faktorer. Til høyre er bare grammatiske og pragmatiske feil uavhengige faktorer.

Feilene som har størst negativ effekt på forståeligheta, er altså leksikalske og til en viss grad grammatiske feil. Setningene tåler derimot langt flere pragmatiske feil uten at det går ut over forståelsen. Heldigvis er det nettopp leksikonet og grammatikken det er mulig å forbedre. Det som er virkelig vanskelig, er å skrive gode regler for de pragmatiske aspekta ved setninga. Samisk har ikke bestemthet som grammatisk kategori og heller ikke obligatoriske determinativer foran adjektiv. I norsk står substantiv i bestemt eller ubestemt form og med bestemt eller ubestemt determinativ avhengig av informasjonsstrukturen i teksten, det vil si om referenten til NPen for eksempel har vært nevnt tidligere i teksten eller

[7] For ei drøfting av CART-metoden, se Strobl m.fl. 2019.

ikke, og om den skal bli forstått generisk, spesifikt eller kontrastivt. Men som evalueringa viser, går slike feil i liten grad ut over forståelsen av de oversatte setningene.

[4.1] *Feil i innputt*

Som vi så ovafor for setning (4), kan feil valg av analyse i forhold til kontekst få katastrofale følger for oversettinga. I tekstene var det også eksempel på at grammatiske feil i originalsetningene blei ført videre i oversettinga. I setninga som begynner med *dát leat boađus das*, ville det vært riktigere med entallsbøyning av kopula: *lea*, og denne feilen gir oversettinga *disse er resultat av* istedenfor *dette er et resultat av ...* Her kunne det vært regler som gjorde valg av numerus i oversettinga avhengig av subjektet *boađus* som er i entall, sjøl om kopula er i flertall. For kopula er denne typen grammatisk feil nokså vanlig i tekster.

[5] KONKLUSJON

Evalueringa av maskinoversettingssystemet viser at det fungerer bedre for tekster på Samisk høgskoles hjemmesider enn for nyhetstekster som krever et større leksikon i oversettingsprogrammet, og også flere regler for leksikalsk seleksjon. Dette kommer dels av at vi har arbeidet spesielt med tekster fra Samisk høgskole, men framfor alt av at nyhetssjangeren er langt mer heterogen, tematisk sett.

Det er helt avgjørende at kildespråkets analyse er korrekt. Viss analysen velger feil ord i setninga som finitt verb, går det ut over analysen, og dermed også oversettinga, av hele setninga.

Det viktigste resultatet av evalueringa var at det er en systematisk forskjell i hvilken innvirkning ulike feiltyper har for forståeligheta. Størst konsekvenser fikk leksikalske feil. Setningene med dårligst skår for forståelse var de som hadde feil ordvalg, særlig der orda ikke var semantisk beslekta, eller hadde innholdsord som ikke var oversatt.

Størst toleranse hadde evaluererne for pragmatiske feil, dvs. feil bruk av bestemthet og manglende eller for mye bruk av determinativer. Dette er språktrekk der valg av den ene eller andre forma ikke nødvendigvis fører til ugrammatiske setninger, men snarere er avhengig av informasjonsstrukturen i teksten.

Det som peker seg ut for å forbedre programmet, er dermed å forbedre ordvalget, særlig av innholdsord, men også av grammatiske ord som preposisjoner og pronomen. For hvert nye domene trengs det også tilpassing, framfor alt av leksikon, og flere regler for leksikalsk seleksjon.

Alt i alt viser evalueringa likevel at det er fullt mulig å få regelbasert maskinoversetting fra samisk til norsk som fungerer godt for forståelse. Best fungerer

det når programmet blir brukt for et kjent domene, i dette tilfellet Samisk høgskoles nettsider. Programmet gir dermed samiskspråklige frihet til å kunne skrive på samisk uten samtidig å måtte oversette teksten sin til norsk.

ETTERORD

Vi kan vanskelig tenke oss en bedre måte å minnes Janne på enn å ved å skrive en artikkel om regelbasert maskinoversetting fra nordsamisk til norsk. Janne har vært sentral i utviklinga av begge analysemetodene som ligger til grunn for arbeidet vårt (FST og CG), den første var temaet for hovedoppgava hennes, og den andre lå til grunn for Oslo-Bergen-taggeren. Gjennom to tiår med arbeid med samisk språkteknologi har vi alltid hatt entusiastisk støtte fra og fine samtaler med Janne og kollegene hennes på Tekstlaboratoriet, fra de første forsøka våre med datastøtta språklæring til arbeidet med det første grammatisk annoterte samiske talespråkskorpuset, LIA. Karakteristisk nok var hun sjøl en av informantene for denne artikkelen. Vi vil savne deg, Janne.

REFERANSER

Antonsen, Lene, Ciprian Gerstenberge, Maja Kappfjell, Sandra Nystø Ráhka, 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*, 123–131. ACL. Gothenburg, Sweden.

Antonsen, Lene & Trond Trosterud. 2017. Ord sett innafra og utafra – en datalingvistisk analyse av nordsamisk. *Norsk Lingvistisk Tidsskrift* 15(1): 153–185.

Bick, Eckhard & Lars Nygaard. 2007. Using Danish as a CG Interlingua. A Wide-Coverage Norwegian-English Machine Translation System. I *Proceedings of the 16th Nordic Conference of Computational Linguistics*. University of Tartu, Estonia.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* (25)2: 127–144.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang & Ming Zhou. 2018: *Achieving Human Parity on Automatic Chinese to English News Translation*.

Johannessen, Janne Bondi. 1990. *Automatisk morfologisk analyse og syntese: Tonivåmodellen benyttet på norsk substantivbøying*. Oslo: Novus forlag.

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

Johnson, Ryan, Tommy Pirinen, Tiina Puolakainen, Trond Trosterud, Francis M. Tyers & Kevin Unhammer. 2017. North-Sámi to Finnish rule-based machine translation system. I *Proceedings of the 21st Nordic Conference on Computational Linguistics,* 115–122. ACL. Gothenburg, Sweden.

Lønning, Jan Tore, Stephan Oepen, Dorothee Beermann, Lars Hellan, John Carroll, Helge Dyvik, Dan Flickinger, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, Victoria Rosén & Erik Velldal. 2004: LOGON. A Norwegian MT effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*, Uppsala, Sweden.

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14: 323–348.

Trosterud, Trond & Kevin Brubeck Unhammer. 2013. Evaluating North Sámi to Norwegian assimilation RBMT. I *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*, vol. 3 of Technical report, edited by Philip Tedeschi & Annie Zaenen, 13–26. Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes & Jeffrey Dean 2016. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.* eprint: arXiv:1609.08144.

SUMMARY

The article presents a rule-based machine translation system from Northern Sami to Norwegian. The grammatical analysis is done with Giellatekno and Divvun's North Sami program for analysis and translation. We have written the transfer component (transfer lexicon and grammatical rules) within the framework of the open machine translation system Apertium. The article contains an evaluation of translated text for two different domains. The translated texts score better on the presentation of the content than on fluent language. By classifying the errors into lexical, grammatical and pragmatic errors, we show that lexical errors are the most harmful for text comprehension. The other two types of errors give a poor language quality, but they have little effect on comprehension. The type of error that is the easiest to correct is the lexical, which is a promising conclusion for the development of a machine translation system for text comprehension.

KONTAKT

Trond Trosterud
UiT Norges arktiske universitet
trond.trosterud@uit.no

Lene Antonsen
UiT Norges arktiske universitet
lene.antonsen@uit.no