OSLa Hagen, Hjelde, Stjernholm & Vangsnes (red.) *Bauta: Janne Bondi Johannessen in memoriam*, Oslo Studies in Language 11(2), 2020. 243–250. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

MOT EN TREBANK FOR AMERIKANORSK

ANDRE KÅSEN
Nasjonalbiblioteket

SAMMENDRAG

I denne artikkelen presenteres en framgangsmåte for å tilordne en del av amerikanorsk talespråkskorpus syntaktiske dependensrelasjoner automatisk. Ulike maskinlæringsteknikker og korpus blir tatt i bruk. Til slutt gis et mål på forventa nøyaktighet og en sammenlikning med en annen relativt nylig publisert trebank for norsk.[1]

[1] INTRODUKSJON

Amerikanorsk er en variant av norsk som tales av etterkommere av nordmenn flere steder på det amerikanske kontinent. Denne artikkelen vil begynne arbeidet mot en trebank for amerikanorsk slik at påstander om amerikanorskens syntaks, det være seg leddstilling eller partikkelplassering, kan undersøkes ved hjelp av presise søk i en større samling med syntaktiske analyser.

En trebank består av manuelt kvalitetsikrede analyser av setninger og fraser innen et bestemt grammatisk-analystisk paradigme. De siste årene har dependensgrammatikk blitt et populært utgangspunkt for å «dyrke» trebanker. Dependensgrammatikken ser helt grunnleggende på syntaktiske funksjoner som relasjoner mellom ord, og dependensgrammatiske analyser blir gjerne uttrykt som grafer. En graf defineres som en mengde noder (ordene) og en mengde kanter (relasjoner mellom ord)[2].

Arbeidet her skal legge til rette for den første amerikanorske trebanken og vil først og fremst anvende maskinlært, automatisk syntaktisk analyse, gjerne kalt *parsing*. Det vil med andre ord ikke føre til en gullstandard-trebank, men til et syntaktisk annotert korpus som vil muliggjøre syntaktiske søk. For at annotasjonene skal bli så presise som mulig, tas det i bruk nyere maskinlæringsteknikker, i tillegg til en rekke relevante ressurser utvikla ved Tekstlaboratoriet i samarbeid med andre.

[1] Takk til Arnstein Hjelde, Øystein A. Vangsnes, Lilja Øvrelid, Marie Samuelsen, Magnus Breder Birkenes, Andrea Myklebust Huus og en anonym fagfelle for tilbakemeldinger.

[2] Nivre (2005) er en god innføring i enkle grafteoretiske forhold, dependensgrammatisk analyse og dependensparsing.

Artikkelen er strukturert som følger: Del 2 presenterer de relevante ressurs ene som er nødvendige for å etablere det syntaktisk annoterte korpuset. I del 3 presenteres maskinlæringsteknikkene som skal brukes med ressursene. I del 4 evalueres kvaliteten på samspillet mellom maskinlæring og ressurser, før det i del 5 oppsummeres.

FIGUR 1: Distribusjon av målepunkter i CANS for statene Nord- og Sør-Dakota, Minnesota, Wisconsin, Iowa og Illinois.

[2] RESSURSER

[2.1] *Amerikanorsk talespråkskorpus*

Amerikanorsk talespråkskorpus (CANS) er gjort rede for i Johannessen (2015). Korpuset inneholder hovedsakelig intervjuer med eldre talere av amerikanorsk, og i tillegg en del samtaler talerene imellom. Informantene er eldre mennesker, ofte i åttiårene, fra USA og Canada med tyngdepunkt i den amerikanske Midtvesten. For delkorpuset vi behandler her, er 168 av 205 informanter og 35 av 45 steder beholdt. Den geografiske fordelingen av de 35 stedene er vist i Figur 1.

Alle opptaka er transkribert lydnært etter standarden man finner i Kåsen et al. (2018) og deretter ordklassetagga med TreeTagger (se Johanesssen 2015 og referanser der). Korpuset er siden utvida til å inneholde amerika-*nordisk* talepråk med blant annet et amerikasvensk delkorpus.

[2.2] *Norsk dependenstrebank*

Norsk dependenstrebank (NDT, se Solberg et al. 2015) er den første dependenstrebanken for norsk, altså en større samling syntaktiske analyser i et dependensgrammatisk rammeverk av norsk (slik vi definerte over). NDT består av bokmål

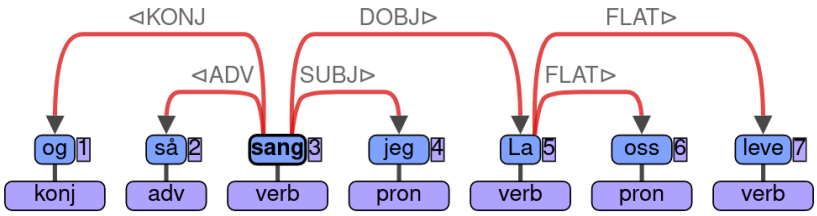
og nynorsk tekst annotert med syntaks, ordklasser og morfologi, og teller 311 000 bokmålseksemplarer og 303 000 nynorskeksemplarer. I Velldal et al. (2017) blei det vist at å bruke både bokmåldelen og nynorskdelen av en trebank er fordelaktig for parsing, derfor vil dette også gjøres i det videre arbeidet her.

For denne arikkelen er det også viktig å merke seg at NDT kun inneholder skriftlige kilder, som nyhetsartikler, stortingstaler, NOU-er og blogginnlegg.

[2.3] *Language Infrastructure made Accesible-korpuset og -trebanken*

LIA-korpuset består av dialektal tale som er lydnært transkribert og translitterert til nynorsk, og korpuset teller i overkant av 2 000 000 eksemplarer. Korpuset inneholder transkripsjoner av informantintervjuer fra hele landet som er blitt samla inn i flere omganger siden 1930-tallet.

En delmengde av LIA-korpuset er annotert manuelt med syntaks, og blir derfor kalt LIA-trebanken. I motsetning til NDT inneholder LIA-trebanken talespråk. Annotasjonen av LIA-trebanken støtter seg på samme retninglinjer som NDT (se Solberg et al. 2015). Likevel er det visse aspekter ved i LIA-materialet som har gjort det nødvendig å utvide annotasjonsstandarden utforma for NDT (Øvrelid et al. 2018). Dette er nødvendig for å analysere såkalte *disfluencies*, altså fenomener hvor talerne retter eller starter nye utsagn før de har gjort seg ferdig med en pågående ytring.



FIGUR 2: Eksempel på en grafisk fremstilling av et segment fra CANS med dependensrelasjoner og ordklassetagger.

[3] METODER

[3.1] *En minimal amerikanorsk trebank*

For å kunne teste resultata av metodene, beskrevet over på amerikanorsk, blei det trukket 25 tilfeldige segmenter fra CANS som er gjennomgått og retta av forfatteren ved hjelp av annoteringsprogrammet Annotatrix[3]. I dette

[3] Verktøyet kan benyttes i nettleseren på følgende url: https://maryszmary.github.io/ud-annotatrix/standalone/annotator.html

annotasjonsarbeidet blei retningslinjene for annotasjon av LIA-trebanken fulgt. I Figur 2 finner man et eksempel på en grafisk framstilling av en av de 25 segmentene.

[3.2] *Dependensparsing*

Dependensgrammatikken analyserer syntaktisk struktur som binære, asymmetriske relasjoner mellom ord. Orda som inngår i en slik relasjon kalles et hode og en dependent. Det er vanlig å si at hodet styrer (eng. *govern*) dependenten. I Figur 2 over så vi et eksempel på en dependensgraf og et segment fra CANS-korpuset som er dependensgrammatisk annotert. Ordet *sang* i Figur 2 ser man ikke pekes på av noen andre ord. Et slikt ord kalles en *rot*, og kun et fåtall dependensrelasjoner kan være såkalte *røtter*. I dette eksempelet er det sjølsagt snakk om et finitt verb.

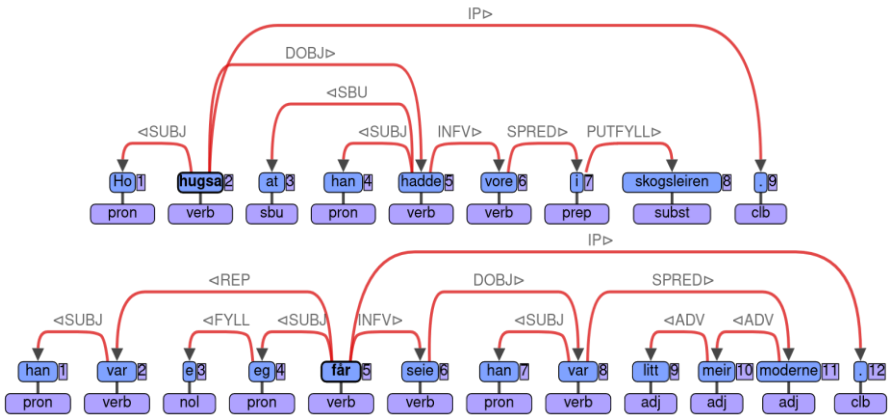
Å automatisere dependensanalyse kalles gjerne dependensparsing. Ved hjelp av ulike algoritmer tilordner man en gitt setning, eller i vårt tilfelle segment, den mest plausible dependensgrafen. For å finne disse dependensgrafene benyttes den syntaktiske parseren UUParser[4]. Denne parseren brukes til å lage tre ulike modeller ved hjelp av de to trebankene vi har tilgjengelig. Dernest skal et delkorpus som består av data fra 36 steder i statene Nord- og Sør-Dakota, Minnesota, Wisconsin, Iowa og Illinois parses. Disse statene utgjør et sammenhenende område som vist i Figur 1. Senere vil man uten store problemer kunne utvide denne prosedyren til resten av målepunkta i korpuset, rette de automatiske analysene og dermed ha en fullgod trebank.

[3.3] *Overført læring*[5]

Maskinlæring har drevet utviklingen innen språkteknologi de siste årene og er en betegnelse på algoritmer som på et vis «lærer» av data. I hovedsak deler man disse inn i *leda* og *uleda* algoritmer eller læring. Mens leda læring avhenger av merkede/annoterte eksempler som i en trebank, er algoritmen i uleda læring kun avhengig av en samling tekst. Når man for eksempel lager en parser – som representerer en form for leda læring – presenterer man parseren for en mengde eksempler. Parseren trenger også å se en fasit eller gullstandardanalyse for det gitte eksempelet. Dette gjentar man et bestemt antall ganger, eller til man har oppnådd et ønska presisjonsnivå.

[4] UUParser er beskrevet i de Lhoneux et al. (2017) og er en transisjonsbasert dependensparser utvikla ved Universitetet i Uppsala. Den er en videreutvikling av BIST-parseren til Kiperwasser & Goldberg (2017).

[5] Terminologien for maskinlæring og kunstig intelligens på norsk er ennå ikke helt etablert, derfor følges terminologien i Teknologirådets rapport om samme emne (https://teknologiradet.no/wp-content/uploads/sites/105/2018/09/Rapport-Kunstig-intelligens-og-maskinlaering-til-nett.pdf) i denne artikkelen.



FIGUR 3: Øverst kan finner man et eksempel fra NDT, mens man underst finner et eksempel fra LIA-trebanken. I det underste eksempelet ser man også et tilfelle av en såkalt *disfluency* (REP).

Om man har få oppmerkede eksempler for liknende oppgaver, kan man benytte seg av såkalte blandingsmodeller, nærmere bestemt overført læring eller transfer learning[6]. Tanken bak slike modeller er at man kan dra nytte av det analytiske rommet de oppmerkede eksemplene deler. Det er for eksempel ingen åpenbar grunn til at analyser av direkte objekter eller liknende i NDT avviker i stor grad fra samme analyser i LIA som man kan se i Figur 3. Fra dette kan vi også [anta at CANS-materialet vil likne disse to trebankene. Slik metoder letter i tillegg behovet for beregningskapasitet en god del og vil presumptivt være et bedre utgangspunkt for videre læring. Og med utgangspunkt i NDT og LIA-trebanken gjøres derfor følgende:

(i) Initialiserer en parsermodell med LIA-trebanken.

(ii) Lager den initielle modellen med NDT.

(iii) Finjusterer så den resulterende NDT-modellen med LIA-trebanken.

Man overfører altså relevante analystiske punkter fra NDT til en modell basert på LIA-trebanken.

[4] EVALUERING

For å få et klarere bilde på utgangspunktet for arbeidet mot en amerikanorsk trebank og presisjonsnivåa til dependensparserene, gjør vi to beregninger: 1) *un-*

[6] En lengre utgreiing om overført læring finner man i Ruder (2019).

labelled attachment score (UAS), som handler om hvor godt parseren klarer å tilordne rett hode til hver dependent (for eksempel at dependenten *jeg* i Figur 2 blir «pekt på» av hodet *sang*) og 2) *label score* (LS), altså i hvor stor grad parseren tilordner rett syntaktisk funksjon eller dependensrelasjon mellom dependenten og hodet (som mellom *sang* og *jeg* i Figur 2 vil tilsvare den syntaktisk funksjonen subjekt eller dependensrelasjonen SUBJ). Disse to beregningene gjør vi ved å la tre ulike modeller, henholdsvis en LIA-modell, en NDT-modell og en overført NDT+LIA-modell, annotere de to ulike delkorpus vi har satt til side for evaluering, nemlig testdelen av LIA-trebanken og vår minimale CANS-trebank.

I Tabell 1 ser vi de aktuelle utregningene. Kolonnen helt til venstre angir hvilken modell som er blitt benyttet, øverste rad angir hvilken evalueringsdata som er blitt benytta og andre rad fra toppen viser hvilken tallfesting de ulike kolonnene angir. Det fremkommer at LIA-trebanken alene gir bedre tall for både LIA- og CANS-evalueringsdataen, mens NDT er betydelig dårligere for CANS enn LIA.

Videre ser vi at når vi *overfører* NDT til LIA og siden evaulerer, gir det oss bedre resultater for både LIA og CANS. Dette er spesielt påfallende for differansen mellom UAS og LS. For både LIA- og NDT-modellen er UAS høyere enn LS for den minimale CANS-trebanken, men når så NDT blandes inn, ser vi at LS faktisk blir høyere enn UAS. Dette kan muligens antyde at det er strukturelle ulikheter mellom de to evalueringsdatasettene, og dette bør undersøkes nærmere.

	LIA		**CANS**	
	UAS	**LS**	**UAS**	**LS**
LIA	83.69%	84.71%	69.23%	66.30%
NDT	71.61%	72.83%	62.64%	61.17%
NDT+LIA	85.28%	86.29%	78.75%	79.12%

TABELL 1: Tre ulike parsermodeller LIA, NDT og NDT+LIA evaluert på LIA-evaueringsdel og den minimale CANS-trebanken. Ytelsen til modellene er tallfesta i *unlabelled attachment score* og *label score* som, henholdsvis, antyder hvor ofte de ulike modellene tilordner en dependent rett hode og hvor ofte en hode-dependent-relasjon merkes med rett syntaktisk funksjon. *Label score* er snittet av alle de ulike mulige funksjonene.

[5] OPPSUMMERING

Ovenfor blei det presentert et første forsøk på å automatisk tilordne syntaktiske relasjoner til amerikanorsk tale. For å nyttegjøre oss av så mange aspekter som mulig av arbeidet som til nå er nedlagt i CANS, samt tidligere arbeider som NDT og LIA, blei det tatt i bruk ulike maskinlæringsalgoritmer.

Til slutt evaluerte og kvantifisterte man forventa kvalitet på trebanken med

en liten håndannotert referansetrebank. En automatisk tilordning av dependensrelasjoner til delkorpuset med talere fra Nord- og Sør-Dakota, Minnesota, Wisconsin, Iowa og Illinois (fordelt som i Figur 1) vil tilgjengliggjøres på et passende vis.

Et fornuftig videre arbeid ville være å manuelt kvalitetsikre et større referansekorpus som vil sette oss i stand til å evaluere mer nøyaktig, men også bruke samme metodikk beskrevet over til å lage en amerikanorsk parser. En nærstudie av de enkelte syntaktiske funksjonene eller dependensrelasjonene er også et mål, i tillegg til å «dyrke» en større amerikanorsk trebank med egne annotasjonsretningslinjer.

DEDIKASJON

Janne ansatte meg som transkribør på Tekstlaboratoriet mellom prosjektene ScanDiaSyn og LIA. Fra Hundremeterskogen i Henrik Wergelands hus på Blindern fikk jeg staka ut en vei i livet i et herlig miljø. Uten en slik mulighet veit jeg ikke hvor jeg ville vært i dag. Takk, Janne!

REFERANSER

Johannessen, Janne B. 2015. The Corpus of American Norwegian Speech (CANS). I *Proceedings of the 20th Nordic Conference of Computational Linguistics*, redigert av Béata Megyesi, 297–300. Linköping University Electronic Press.

Johannessen, Janne B., Kristin Hagen, Live Håberg, Signe Laake, Åshild Søfteland & Øystein A. Vangsnes. 2009. *Transkripsjonsrettleiing for ScanDiaSyn. Teknisk rapport.* Tekstlaboratoriet, Universitetet i Oslo.

Kiperwasser, Eliyahu & Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4: 313–327.

Kåsen, Andre, Eirik Olsen, Sjånes Rødvand, Linn Iren & Eirik Tengesdal. 2018. *Transkripsjons- og translittereringsveiledning for Norsk i Amerika.* Teknisk rapport. Tekstlaboratoriet, Universitetet i Oslo.

de Lhoneux, Miryam, Sara Stymne & Joakim Nivre. 2017. Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. I *Proceedings of the 15th International Conference on Parsing Technologies*, redigert av Y. Miyao & K. Sagae, 99–104. Association for Computational Linguistics.

Nivre, Joakim. 2005. Dependency Grammar and Dependency Parsing. Teknisk rapport (MSI report 05133). School of Mathematics and Systems Engineering, Växjö University

Ruder, Sebastian. 2019. Neural transfer learning for natural language processing. Doktorgradavhandling. National University of Ireland, Galoway.

Solberg, Per Erik, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen & Janne B. Johannessen. 2014. The Norwegian Dependency Treebank. I *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, redigert av Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis, 789–795. European Language Resources Association.

Velldal, Erik, Lilja Øvrelid & Petter Hohle. 2017. Joint UD Parsing of Norwegian Bokmål and Nynorsk. I *Proceedings of the 21st Nordic Conference on Computational Linguistics*, redigert av Jörg Tiedemann & Nina Tahmasebi, 1–10. Association for Computational Linguistics.

Øvrelid, Lilja, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per E. Solberg & Janne B. Johannessen. 2018. *The LIA treebank of spoken Norwegian dialects.* I *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, redigert av Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani & Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Takenobu Tokunaga, 4482–4488. European Language Resources Association.

SUMMARY

This article presents a method for automatic assignment of syntactic dependency relations to the corpus of American Norwegian speech (CANS). Different machine learning techniques and corpora are used. Finally, an accuracy measure is computed and compared with a relatively new treebank for spoken Norwegian.

KONTAKT

Andre Kåsen
Nasjonalbiblioteket
andre.kasen@nb.no