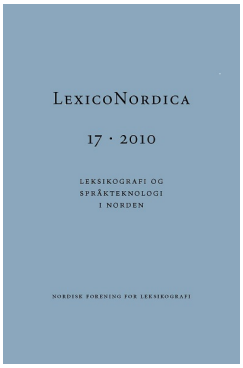
# LexicoNordica

| | |
|---|---|
| Titel: | Bruk av et norsk leksikon til tagging og andre språkteknologiske formål |
| Forfatter: | Kristin Hagen og Anders Nøklestad |
| Kilde: | LexicoNordica 17, 2010, s. 55-72 |
| 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.

# Bruk av et norsk leksikon til tagging og andre språkteknologiske formål

*Kristin Hagen & Anders Nøklestad*

*Norsk ordbank* (the Norwegian Word Bank) is an electronic lexicon for the two Norwegian written standards, *Bokmål* and *Nynorsk*. It forms the basis of many, probably most, of the existing language technology tools for Norwegian. The lexicon is based on the entries and inflectional information found in the dictionaries *Bokmålsordboka* and *Nynorskordboka* as well as word lists and inflectional patterns developed by IBM Norway. We present some background information about the lexicon and show how it has been applied to a variety of language technology tools and various applications for end users. Since the lexicon was developed from resources meant for use by human readers, much work has been devoted to modifying the lexicon to make it better suited for use in language technology, and the main focus of our paper is on this work.

## 1. Innledning

Norsk ordbank er et elektronisk leksikon for de to norske skriftspråksstandardene bokmål og nynorsk. Innholdet i Ordbanken kommer dels fra håndordbøkene *Bokmålsordboka* og *Nynorskordboka*, og dels fra elektroniske ordlister utviklet ved IBM Norge. Ordbanken ble laget i 1996 i forbindelse med Taggerprosjektet, et prosjekt som skulle utvikle en morfologisk og syntaktisk tagger for bokmål og nynorsk.

Mye av bakgrunnsmaterialet til Ordbanken stammer altså fra ordbøker, skrevet og redigert for å tilfredsstille menneskelige lesere i et trykt medium. I den første delen av denne artikkelen vil vi redegjøre for opprinnelsen til Ordbanken med fokus på hvilke endringer og tilpasninger som måtte gjøres for å få innholdet best

mulig egnet som elektronisk ordbok brukt til morfologisk og syntaktisk tagging.

Norsk ordbank er et enkelt fullformsleksikon som mangler semantisk tilleggsinformasjon som ville ha vært nyttig i en del sammenhenger. Etter Taggerprosjektets avslutning har Ordbanken likevel vært brukt i mange ulike språkteknologiske applikasjoner og verktøy. Den siste delen av artikkelen vil gå nærmere inn på dette og beskrive noen av verktøyene i større detalj.

## 2. Kort om bakgrunnen til Norsk ordbank

Norsk ordbank, eller *taggerbasen* som leksikonet ble kalt i begynnelsen, ble opprettet i 1996 i regi av *Taggerprosjektet*.[1] Formålet med dette prosjektet var å utvikle en disambiguerende morfologisk og syntaktisk tagger for norsk,[2] og for å komme i gang med dette arbeidet trengtes det elektroniske fullformsordlister. Ved prosjektstart fantes det ikke andre lister tilgjengelig enn IBM-ordlistene som *Dokumentasjonsprosjektet* hadde kjøpt fra *IBM Norge*.[3] For å komme i gang med utviklingen av taggeren ble det derfor beslut-

1 Taggerprosjektet (1996–1999, ledet av Janne Bondi Johannessen) ble beregnet til seks årsverk. Tre av disse ble finansiert fra Norges forskningsråd, to fra Dokumentasjonsprosjektet (ledet av Christian-Emil Smith Ore) og ett fra Tekstlaboratoriet (ledet av J. B. Johannessen). I tillegg kom grunnmateriale fra *Bokmålsordboka* og *Nynorskordboka* utviklet ved seksjon for Leksikografi ved Universitetet i Oslo, IBM-ordlistene med grammatiske koder for bokmål og nynorsk, programvare for disambigueringsdelen av taggeren fra Lingsoft og argumentstruktur for verb fra NorKompLeks, NTNU.

2 Taggeren som ble utviklet, heter i dag Oslo-Bergen-taggeren, se referanse.

3 Samtidig med Taggerprosjektet (1996) fikk NorKompLeks-prosjektet ved NTNU støtte fra Norges forskningsråd til å lage et maskinleselig språkteknologisk leksikon for norsk (bokmål og nynorsk). Dette arbeidet kunne Taggerprosjektet dessverre ikke benytte seg av, siden prosjektet var avhengig av elektroniske ordlister fra prosjektstart for å utvikle taggeren.

tet å bruke disse ordlistene sammen med oppslagsord og grammatiske opplysninger fra *Bokmålsordboka* og *Nynorskordboka* for å lage en ordbase med fullformer for taggeren. Dette arbeidet blir beskrevet i større detalj i kapittel 3.

Flere miljøer ved Universitetet i Oslo (UiO) var involvert i utviklingen av taggerbasen: Dokumentasjonsprosjektet, Seksjon for leksikografi og målføregransking ved Institutt for nordistikk og litteraturvitenskap og Tekstlaboratoriet ved Institutt for lingvistiske fag. For å sikre basens videre utvikling, vedlikehold og drift fikk taggerbasen i 2001 et styre med overordnet ansvar for dette. Styret har medlemmer fra de ulike UiO-miljøene samt Språkrådet. Basen ble omdøpt til *Norsk ordbank* og eies i dag av Institutt for lingvistiske og nordiske studier (ILN), der alle miljøene nevnt ovenfor nå er samorganisert, og Språkrådet.

Ordbanken blir finansiert ved hjelp av egeninnsats fra ILN og Språkrådet samt med enkelte midler fra salg. Ordbanken er for øvrig nedlastbar på GPL-lisens.[4]

Det har bare vært gjort mindre endringer i Ordbanken siden oppstarten i 1996. LOGON-prosjektet (Oepen et al. 2007) finansierte noe arbeid for bokmål i 2001, og Språkrådet finansierte oppdateringer etter nyere Språkråds-vedtak i 2007, samt innlegging av nyord fra Bokmålsordboka og Nynorskordboka.

Ordbanken inneholder i dag 151 229 lemmaer for bokmål og 126 323 lemmaer for nynorsk.

## 3. Utviklingen av et elektronisk fullformsleksikon for norsk

Utviklingen av Ordbanken var fra starten av styrt av behovet til taggeren som skulle utvikles. For å gjenkjenne alle ord i en gitt

4 GNU General Public License (GPL), jf. http://www.gnu.org/licenses/gpl.html

tekst trengte taggeren fullformsordlister, det vil si ordlister med lemmaer (for eksempel *bil*) og alle lemmaets mulige bøyingsformer (for eksempel *bil bilen biler bilene*). Det var i tillegg ønskelig å lagre lemmaene og deres bøyingsformer på en måte som gjorde at de var lette å finne igjen og lette å redigere om det skulle vedtas endringer i ortografi eller bøyingsmåte senere. Det var også ønskelig at bøyingskodene skulle følge ordklasseterminologien i *Norsk referansegrammatikk (*Faarlund/Lie/Vannebo 1997) med for eksempel *determinativer* i stedet for *artikler*, og *subjunksjoner* i stedet for *underordnende konjunksjoner.*

I kapitlene nedenfor vil vi gå igjennom de ulike kildene til Ordbanken og beskrive nærmere hva som ble gjort.

## 3.1. IBM-ordlistene

Listene fra IBM (Engh 1994) inneholdt nærmere 122 000 lemmaer for bokmål og mer enn 110 000 for nynorsk. Lemmaene var hentet fra både korpus og ordbøker, og alle ord var utstyrt med en bøyingskode slik at fullformene kunne avledes, se eksempel 1 på neste side.

Listene ble utviklet for å brukes i IBMs språkteknologiske produkter, for eksempel i IBMs stavekontroll. Når listene nå skulle brukes til tagging, hadde de en del åpenbare svakheter:

1. Listene inneholdt ikke sideformer.[5]
2. Listene var normative.
3. Bøyingsmønstrene var komplekse og krevende å vedlikeholde.

En tagger bør kunne gjenkjenne alle ordene i en tekst, også sideformer og de mest vanlige feilskrevne ord og bøyinger. IBM-listene måtte derfor suppleres, både med hensyn til lemmautvalg og med hensyn til bøyingsinformasjon, se avsnitt 3.2. og 3.3.

5 *Sideform* er definert slik i *Bokmålsordboka*: "ord- el. bøyningsform i offisiell rettskrivning som ikke kan brukes i lærebøker og i sentraladministrasjonen, t forskj fra *hovedform*".

```
#B_N10075.010
*subst fem
**fullst

Ø900 Ø  NOB_FN
01
02 a,en
03 er
04 ene
05 s
06 as,ens
07 ers
08 enes
```

Eksempel 1: Bøyingskode for bokmålssubstantiv som for eksempel *dør*. Ved hjelp av bøyingsmønsteret kan fullformene *dør, døra* eller *døren, dører, dørene* avledes sammen med tilhørende genitivsformer *dørs, døras* eller *dørens, dørers, dørenes.*

Bøyingsmønstrene fra IBM måtte også endres av samme årsak. Fordi IBM-listene skulle brukes normativt, var det lagt mye arbeid i å normere bruken av flertall for substantiv, komparative former for adjektiv og adjektivavledninger av verb. Dette arbeidet ble gjort i samarbeid med Språkrådet. For taggeren, derimot, var denne normeringen mer til skade enn til gagn. Flertallssubstantiv som *adganger* og *afasier* er ikke vanskelig å finne ved søk i korpus, selv om ordene har en bøyingskode som utelukker flertallsformer i IBM-listene. Og igjen, uansett hva man måtte mene om flertallsformer av slike substantiv, må taggeren kunne tagge eksempler som ”Dette gjelder også for *adganger* som er gitt til grupper” eller ”en gruppe språkforstyrrelser som kalles *afasier*”. Derfor ble alle substantiv – med unntak av ubøyelige substantiv som *gøy* – gitt flertallsformer. Alle adjektiver ble også gitt komparative former

og alle verb fikk adjektivavledninger slik at eksempler som "... det generelt er en myte at noen språk liksom er *ordfattigere* enn andre" og "Jeg syntes de lager for *brummete* lyder" kunne få en analyse av taggeren. (*Ordfattig* hadde opprinnelig kode uten komparative former, og *brumme* hadde kode uten adjektivavledning.)

Som eksempel 1 ovenfor viser, var bøyingskodene i IBM-listene komplekse og uoversiktlige med flere mulige bøyinger for hver kode. Dette ble løst opp slik at systemet skulle bli lettere å forstå og vedlikeholde. I Ordbanken ble f.eks. koden for ordet *dør* fra eksempel 1 løst opp, slik at ordet nå har to bøyingskoder, 700 og 900, som genererer de samme fullformene som tidligere,[6] se eksempel 2. Legg merke til at hvert bøyingssuffiks er unikt definert med et tall som angir grammatisk informasjon for dette suffikset (f.eks. angir kombinasjonen 700, 02 bestemt form entall av substantiv).

| 700 | 900 |
|---|---|
| ---- | ---- |
| 01 | 01 |
| 02 en | 02 a |
| 03 er | 03 er |
| 04 ene | 04 ene |

Eksempel 2: Bøyingskoder for bokmålssubstantiv som for eksempel *dør*. Ved hjelp av bøyingsmønstrene 700 og 900 kan fullformene *dør, døra* eller *døren, dører, dørene* avledes.

IBM-listene inneholdt ikke bare de vanligste lemmaene, men også en god del egennavn og mange kreative sammensetninger av typen *bruksvakthold* og *prisdepartement*. Slike sammensetninger er kanskje

6 I Ordbanken er genitivs-*s* ikke med i bøyingskodene. Oslo-Bergen-taggeren har en egen modul som forsøker å avgjøre om en apostrof eller en *s* markerer gentitiv. Grunnen til dette er at genitiv ikke er en bøyingskategori ved norske substantiver, men at den såkalte genitivs-*s* i stedet er et klitikon som kan hektes på en frase: "jenta som roptes (genitivs-*s*) hatt".

ikke så frekvente, men gjør heller ingen skade for taggeren. Oslo-Bergen-taggeren har for øvrig en egen sammensetningsmodul som analyserer nylagede sammensetninger, slik at en ikke er avhengig av en fullformsordliste med alle mulige tenkelige sammensetninger.

## 3.2. Innhold fra *Bokmålsordboka* og *Nynorskordboka*

Fra *Bokmålsordboka* og *Nynorskordboka* fikk Ordbanken oppslagsord med grammatiske opplysninger og normeringsinformasjon. Lemmatilfanget i ordbøkene var bare delvis overlappende med lemmatilfanget i IBM-ordlistene, i tillegg til at ordbøkene naturligvis inneholdt alle sideformer. Med ordbøkene ble Ordbanken altså styrket med flere lemmaer og mer informasjon om hvert lemma, men også her oppstod det problemer i forhold til taggerens behov:

1. De grammatiske opplysningene fra ordbøkene var lite spesifikke.
2. Ordbøkene hadde flertydige faste uttrykk som oppslagsord.
3. Ordbøkene inneholdt mange lavfrekvente oppslag med svært frekvente homonymer.

Ordbøkenes bøyingskoder, for eksempel *m1*, er ikke spesifikke nok til å kunne konverteres automatisk til bøyingskodesystemet vi endte opp med å bruke i Ordbanken. I Nynorskordboka har for eksempel både *gut* og *låve* bøyingskode *m1*, selv om låve ender på *-e* og dermed trenger en egen bøyingskode *702*. (Med bøyingskode *700* som *gut* har, ville bøyingen for *låven* blitt *låveen*.) For å gi alle ordene i Ordbanken bøyingskoder slik som ordene fra IBM-listene, ble det forsøksvis laget konverteringslister (*m1* → *700*, *f1* → *900*, *v1* → *001* osv.), før lemmaer med bøyingskoder ble korrekturlest manuelt.

For å gjøre ordbøkene mer leservennlige, er faste uttrykk som *av garde* og *til syne* gjengitt som oppslagsord. Dette er oftest en

fordel for taggeren, fordi ordene i uttrykkene hører så fast sammen og bør analyseres sammen. Flertydige faste uttrykk som dette er imidlertid et problem: "Han ville gjøre det *selv om* han ikke fikk lov". Dersom *selv om* er et fast uttrykk, vil taggeren aldri få anledning til å analysere bruken der *selv* og *om* utgjør egne ledd, som i "Hun vasket seg *selv om* kvelden". Uttrykkene fra ordbøkene ble derfor gjennomgått manuelt, og flertydige uttrykk ble merket på en slik måte at de blir utelatt fra taggerens analyse.

For taggeren er det som regel en stor fordel at leksikonet er så rikholdig som mulig. Men når frekvente lemmaer som *kan* (verb), *med* (preposisjon) og *bare* (adverb) har svært lite frekvente homonymer som gjengitt i eksempel 3 nedenfor, skaper det unødig flertydighet for taggeren.

*kan*: (substantiv) prinsetittel i Mongolia[7]
*med*: (substantiv) siktemerke; formål, mening
*å bare seg*: (verb) avholde seg

Eksempel 3: lite frekvente lemmaer som har frekvente homonymer

Gjennom arbeidet med taggeren ble det oppdaget flere slike homonymer, som ble merket slik at de ikke kommer med i taggerens analyse.

Selv om ikke grammatikken fra ordbøkene kunne brukes direkte, er den grammatiske informasjonen tatt inn i Ordbanken sammen med normeringsopplysninger. Etymologi og definisjoner er ikke med, men oppslagsordenes referansenummer er tatt vare på, slik at ordbank og ordbøker kan lenkes sammen.

7 *Kan* og *khan* var sidestilt i *Bokmålsordboka* da Ordbanken ble lagd, men fra 2005 står bare *khan* oppført.

## 3.3. Tillegg

Mange tekster er dårlig korrekturlest og inneholder feil. Feilstavinger og feilbøyinger hører vanligvis ikke hjemme i ordbaser, men siden Ordbanken ble utviklet for tagging, ble det likevel lagt til en del frekvente feil. Slike tillegg fikk *tillegg* oppført som kilde og *unormert* som normering. Nedenfor følger noen eksempler på feil som ble lagt inn i Ordbanken:

- Frekvente feilstavinger av lemmaer: *almen* i stedet for normert *allmenn* (adjektiv), *arbeide* i stedet for normert *arbeid* (substantiv)
- Frekvente feilbøyinger: *gutta* i stedet for normert *guttene, lederer* i stedet for normert *ledere*

Det ble også lagt inn en del nye lemmaer, forkortelser og uttrykk:

- Nye ord/slang/engelske ord: *drita*, *body*
- Flere egennavn
- Uttrykk: *dann og vann*
- Forkortelser: *adj*

I ordbøkene kan enkelte oppslagsord ha betydninger med ulik ordklasse. *All* er for eksempel oppgitt med ordklasse pronomen i Bokmålsordboka, men i betydningsnummer 5 i ordartikkelen står det likevel at ordet kan være adverb i formen *alt*:

> 5 adv i formen *alt*: helt, allerede *alt fra hun var født* / *han er alt reist* / *alt det hun klager* samme hvor mye hun klager / *alt etter forholdene* i samsvar med / *det er alt etter som en tar det* det kommer an på hvordan ...

*Alt* er derfor lagt inn som adverb som *tillegg*.

Tilleggene er lagt inn undervegs i forbindelse med arbeidet med taggeren. Alle tillegg stammer fra korpus, men det er ikke systematisk registrert feil og nyord i Ordbanken.

## 3.4. Argumentstruktur for verb fra NorKompLeks

*NorKompLeks*-prosjektet (Nordgård 1996) utviklet maskinleselige leksikografiske produkt parallelt med arbeidet med taggeren. Da prosjektet ble avsluttet, ble opplysninger om verbenes argumentstruktur lagt inn i Ordbanken på denne måten:

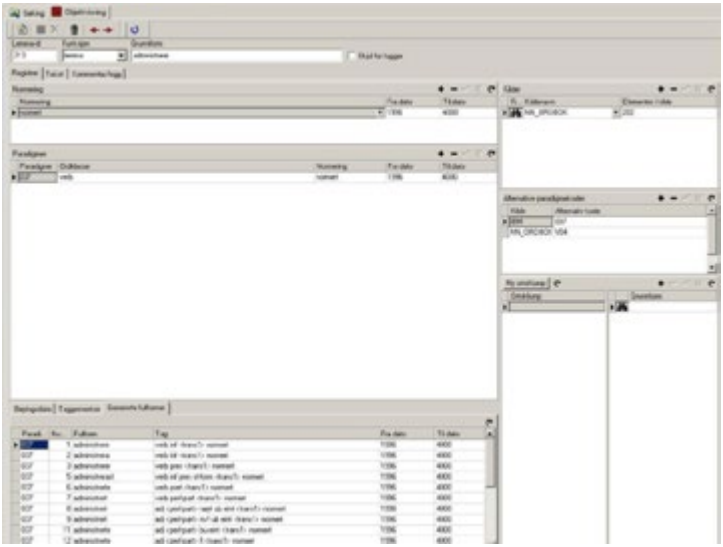
arbeide <intrans1> <predik13>
øke <trans1> <intrans2> <part4/på>

Subkategoriseringsinformasjonen er ikke vedlikeholdt etter prosjektslutt, og for taggerens del viste det seg at informasjonen ikke er fullstendig nok til å kunne brukes i taggeren.

## 3.5. Database og redigeringsverktøy

EDD (Enhet for digital dokumentasjon[8]) ved ILN, Universitetet i Oslo, er ansvarlig for databasedesign, utvikling og teknisk drift av Ordbanken i dag. Innholdet i Ordbanken er lagret i en Oracle-database, og har et grensesnitt for redigering som også er utviklet av EDD, se figur 1.

8 Enhet for digital dokumentasjon (EDD) ble opprettet for å vedlikeholde og videreutvikle databasene og de elektroniske samlingene fra Dokumentasjonsprosjektet.



Figur 1: Utsnitt av Ordbankens grensesnitt for redigering

## 4. Eksempler på bruk av Ordbanken

Som nevnt tidligere inneholder Ordbanken relativt enkel informasjon, men denne informasjonen er til gjengjeld svært grunnleggende og nyttig for mange formål, både i forbindelse med språkteknologiske verktøy og for mer brukerrettede applikasjoner. Noen eksempler på kommersiell bruk av Ordbanken er:

- maskinoversettelse
- søkemotorer
- stavekontroll
- ordspill (nettbaserte spill som involverer ord, som f.eks. Scrabble-liknende spill)
- pedagogiske verktøy

Ordbanken har også blitt brukt i mange – trolig de fleste – ikke-kommersielle prosjekter for utvikling av norsk språkteknologi, særlig i kraft av å fungere som leksikon i Oslo-Bergen-taggeren. Noen eksempler på ikke-kommersielle anvendelser av Ordbanken er:

- ordbokssøk på nett[9]
- de fleste norske korpus ved universitetene i Oslo og Bergen
- talespråkstaggere
  - NoTa-taggeren (Nøklestad/Søfteland 2007; Søfteland/Nøklestad 2008)
  - nynorsk dialekttagger (utviklet i forbindelse med prosjektet *Dialektendringsprosessar* ved Universitetet i Bergen)
- maskinoversettelse (LOGON, Apertium)
- analyse av setninger brukt i trebank og grammatikkspill på nett (VISL; Bick 2005)

For ytterligere å demonstrere nytteverdien av Ordbanken, vil vi nå gå litt mer i detalj om to verktøy som nylig har blitt utviklet for bokmål, og som gjør god bruk av informasjonen i Ordbanken: en navnetypegjenkjenner og et anaforløsingssystem. Begge disse verktøyene er beskrevet nærmere i Nøklestad (2009).

## 4.1. Navnetypegjenkjenning

Navnetypegjenkjenning (eng. *named entity recognition*) går ut på å klassifisere såkalte *named entities*, først og fremst egennavn, i henhold til et sett av kategorier.[10] Nøklestad (2009) beskriver en navnetypegjenkjenner for bokmål som ble utviklet i forbindelse med

9 http://www.bokmålsordboka.uio.no
http://www.nynorskordboka.uio.no

10 Mange navnetypegjenkjennere klassifiserer også datoer, tidsuttrykk, prosenter og pengebeløp. Systemet som er beskrevet her, fokuserer imidlertid på egennavn.

Nomen Nescio-prosjektet (Johannessen et al. 2005), et prosjekt som hadde som formål å lage navnetypegjenkjennere for norsk, svensk og dansk. Prosjektet opererte med følgende navnekategorier:

- Person
- Organisasjon
- Sted
- Hendelse (f.eks. *Statens høstutstilling*, *Kristiansand box cup*)
- Verk (bøker, filmer, musikkalbum o.l.)
- Annet (f.eks. *Bovine-virus*, *Nissan Terrano*)

Systemet som er beskrevet av Nøklestad, bruker såkalte maskinlæringsteknikker for å klassifisere hvert egennavn i en tekst i henhold til disse kategoriene. Maskinlæringsteknikker er teknikker som setter en datamaskin i stand til å lære å utføre en oppgave (f.eks. å klassifisere navn) ved å se på et stort antall riktig klassifiserte eksempler i stedet for at den blir gitt et sett av regler for hvordan oppgaven skal løses. For at en maskin skal kunne lære å klassifisere egennavn, må den få informasjon om hvert navn og omgivelsene navnet står i. Nøklestads system har tilgang til følgende informasjon:

- formen på navnet
  - ”suffikser” (dvs. de tre siste bokstavene i ordet)
  - antall ord i navnet
  - hvorvidt navnet bare inneholder store bokstaver (f.eks. IBM)
- lemmaet til navnet og andre ord i nær kontekst
- ordklassen til ordene i nær kontekst
- syntaktiske relasjoner
- navnelister

Viktige deler av denne informasjonen blir hentet fra Norsk ordbank. Teksten blir disambiguert av Oslo-Bergen-taggeren, som henter lemmaer og ordklasser direkte fra Ordbanken. De syntaktiske relasjonene blir også bestemt av Oslo-Bergen-taggeren, og siden Ordbanken utgjør taggerens leksikon, bidrar den også indirekte til denne typen informasjon. Navnetypegjenkjenneren oppnår et prestasjonsnivå på 83 %, noe som er et godt resultat i internasjonal sammenheng. Dette viser at den enkle informasjonen som finnes i Ordbanken, kombinert med navnelister og ordformer, er tilstrekkelig for å oppnå et godt resultat på denne oppgaven; semantisk informasjon er altså ikke en forutsetning.

## 4.2. Anaforløsing

Hovedtemaet i Nøklestad (2009) er utviklingen av et system for automatisk anaforløsing (eng. *anaphora resolution*) i bokmål. Anaforløsing, slik denne oppgaven er definert hos Nøklestad, innebærer å finne den nærmeste antecedenten for hver anafor i en tekst.[11] Dette er illustrert i eksempel 4 nedenfor, der oppgaven går ut på å finne den nærmeste antecedenten (om noen) for anaforen *det* i hvert enkelt tilfelle.

*Toget traff reinsdyret fordi ...*
a) ... *det kjørte for fort*
b) ... *det sto i skinnegangen*
c) ... *det var mørkt*

Eksempel 4: I a) refererer *det* til toget, og *Toget* er derfor nærmeste antecedent; i b) er det *reinsdyret* som er nærmeste antecedent, mens i c) er *det* ikke-referensielt og har derfor ingen antecedent.

11 Nøklestads system er begrenset til å håndtere pronominale anaforer.

I likhet med navnetypegjenkjenneren som ble beskrevet i forrige underkapittel, baserer også anaforløsingssystemet seg på maskinlæring. For å velge riktig antecedent trenger systemet informasjon om anaforen, potensielle antecedenter og forholdet mellom anafor og potensiell antecedent. Nøklestads anaforløsingssystem benytter seg av følgende informasjonskilder:

- avstand mellom anafor og antecedentkandidat
- ordform
- lemma
- ordklasse
- morfosyntaktiske trekk som genus og numerus
- syntaktiske funksjoner
- semantisk informasjon

Igjen kommer vesentlige deler av denne informasjonen fra Norsk ordbank. Lemma, ordklasse og morfosyntaktiske trekk blir hentet direkte fra Ordbanken via Oslo-Bergen-taggeren. Ordklasse er helt essensielt, siden bare nominale ledd blir ansett som potensielle antecedenter i dette systemet. Lemma og morfosyntaktiske trekk (genus og numerus) viser seg også å være blant de aller viktigste informasjonstypene (Nøklestad 2009:216, 226).

Når det gjelder semantisk informasjon, så utgjør navnetypen på antecedentkandidater som er egennavn den klart viktigste faktoren (Nøklestad 2009:217, 226). Navnetypen blir avgjort av navnetypegjenkjenneren beskrevet i forrige underkapittel, og denne navnetypegjenkjenneren benytter seg som nevnt i stor grad av informasjon fra Ordbanken. Som vi også nevnte i forrige underkapittel, bidrar Ordbanken indirekte til disambigueringen av syntaktiske funksjoner, siden den utgjør leksikonet til Oslo-Bergen-taggeren. Vi kan derfor konkludere med at Ordbanken er involvert i de fleste og de viktigste faktorene som systemet bygger på. Dermed spiller den – til tross for sitt relativt enkle innhold – en

avgjørende rolle for å bringe anaforløsingssystemet opp til et prestasjonsnivå på 74,6 %, noe som er et godt resultat i internasjonal sammenheng og signifikant bedre enn det som tidligere har blitt oppnådd for anaforløsing i norsk.

## 5. Konklusjon

I denne artikkelen har vi vist hvordan leksikonet Norsk ordbank ble til, og hvordan det måtte bearbeides for å kunne brukes til grammatisk tagging og andre språkteknologiske formål. Vi har også gitt noen eksempler på hvordan et relativt enkelt leksikon kan utgjøre grunnlaget for komplekse språkteknologiske verktøy som oppnår gode resultater.

## Litteratur

### Ordbøker

Bokmålsordboka = Wangensteen, Boye (red.) 2004: *Bokmålsordboka: definisjons- og rettskrivningsordbok*. Oslo: Kunnskapsforlaget.

Nynorskordboka = Anne Engø/Marit Hovdenak/Dagfinn Worren (red.) 2006: *Nynorskordboka. Definisjons- og rettskrivingsordbok*. Oslo: Det Norske Samlaget.

### Annen litteratur

Bick, Eckhard 2005: Grammar for Fun: IT-based Grammar Learning with VISL. I: Peter Juel Henrichsen (red.): *CALL for the Nordic Languages*. København: Samfundslitteratur (Copenhagen Studies in Language), 49–64.

Faarlund, Jan Terje/Lie, Svein/Vannebo, Kjell Ivar 1997: *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Johannessen, Janne Bondi/Hagen, Kristin/Haaland, Åsne/Jónsdottir, Andra Björk/Nøklestad, Anders/Kokkinakis, Dimitris/Meurer, Paul/Bick, Eckhard/Haltrup, Dorte 2005: Named Entity Recognition for the Mainland Scandinavian Languages. I: *Literary and Linguistic Computing 20(1)*, 91–102.

Nordgård, Torbjørn 1996: NorKompLeks: Some Linguistic Specifications and Applications. *ALLC-ACH '96*. Bergen, 214–216.

Nøklestad, Anders 2009: *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection.* PhD thesis, University of Oslo. Oslo: Acta Humaniora.

Nøklestad, Anders/Søfteland, Åshild 2007: Tagging a Norwegian Speech Corpus. I: Joakim Nivre/Heiki-Jaan Kaalep/Kadri Muischnek/Mare Koit (red.): *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, 245–248.

Oepen, Stephan/Velldal, Erik/Lønning, Jan Tore/Meurer, Paul/Rosén, Victoria/Flickinger, Dan 2007: Towards Hybrid Quality-Oriented Machine Translation. On Linguistics and Probabilities in MT. I: *The 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, 144–153.

Søfteland, Åshild/Nøklestad, Anders 2008: Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. I: Janne Bondi Johannessen/Kristin Hagen (red.): *Språk i Oslo. Ny forskning omkring talespråk*. Oslo: Novus forlag, 226–234.

## Internetthenvisninger

Apertium = http://www.apertium.org/

Dokumentasjonsprosjektet = http://www.dokpro.uio.no/

EDD = http://www.edd.uio.no/index.html

Engh, Jan 1994. IBMMORF: IBM Norges leksikon og morfologi for moderne norsk. Dokumentasjonsprosjektet, Universitetet i Oslo, http://www.uio.no/~janengh/IBMmorf.htm.

Lingsoft = http://www.lingsoft.fi/

LOGON = http://www.emmtee.net/index.php?page=1&lang=no

Norsk ordbank = http://www.edd.uio.no/prosjekt/ordbanken/

Taggerprosjektet – Oslo-Bergen-taggeren = http://www.hf.uio.no/tekstlab/tagger.html

Tekstlaboratoriet = http://www.hf.uio.no/tekstlab/index.html

NorKompLeks = http://www.hf.ntnu.no/hf/prosjekter/spraktek/prosjekter/nkl

VISL = http://visl.sdu.dk/

Kristin Hagen
språkingeniør
Tekstlaboratoriet
Institutt for lingvistiske og nordiske studier
Universitetet i Oslo
Postboks 1102 Blindern
NO-0317 Oslo
kristin.hagen@iln.uio.no

Anders Nøklestad
språkingeniør, ph.d.
Tekstlaboratoriet
Institutt for lingvistiske og nordiske studier
Universitetet i Oslo
Postboks 1102 Blindern
NO-0317 Oslo
anders.noklestad@iln.uio.no