NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Språkteknologiens behov for leksikalsk informasjon

Forfatter: Victoria Rosén

Kilde: Nordiske Studier i Leksikografi 12, 2013, s. 13-41
Rapport fra Konferanse om leksikografi i Norden, Oslo 13.-16. august 2013

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

Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) 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.

Språkteknologiens behov for leksikalsk informasjon

Victoria Rosén

The information to be included in a lexicon depends on its intended use. Automatic syntactic analysis of a large and varied text corpus requires detailed information on inter alia countability, compounds, spelling and inflectional variants, neologisms, and multiword expressions, information that cannot always be harvested from traditional dictionaries. In this article it is shown how deep syntactic analysis of a corpus can contribute to the enrichment of lexical resources.

1. Introduksjon

Språkteknologiske applikasjoner trenger detaljert leksikografisk informasjon om flest mulig ord. Mange slike applikasjoner bruker leksikalske ressurser som har vært utviklet som tradisjonelle papirordbøker, der hensikten har vært å gi mennesker nyttige opplysninger om bruken av ordene i språket. Datamaskinelle programmer kan ha behov for andre typer leksikalsk informasjon, eller for at informasjonen organiseres på en annen måte.

Den leksikalske informasjonen som er nødvendig i oppbyggingen av en trebank for norsk, kan illustrere dette

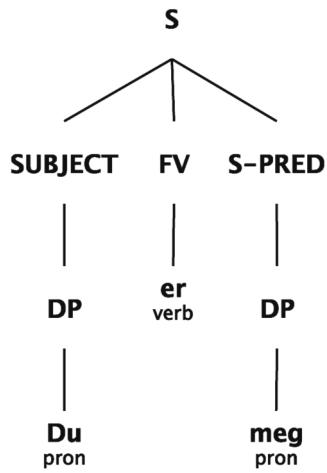
behovet. INESS lager en forskningsinfrastruktur for trebanker, som er korpora annotert med syntaktiske strukturer.[1] Prosjektet har to hovedmål: å lage en infrastruktur for trebanker og å lage en stor trebank for norsk. Det å lage en infrastruktur innebærer å gjøre trebanker lettilgjengelige for brukere; det skal være mulig å finne trebankene online og søke i dem gjennom å bruke en vanlig nettleser. Hovedtrekkene i hva trebanker er, og hvordan den norske trebanken i INESS utvikles, blir presentert i avsnitt 2 nedenfor. I avsnitt 3 vises det hvordan informasjon fra et språkteknologisk leksikon[2] som er utviklet fra en tradisjonell ordbok, kan suppleres og justeres slik at den syntaktiske analysen i trebanken blir mer treffsikker. Avsnitt 4 viser hvordan ord som ikke finnes i tradisjonelle ordbøker, og som man kanskje ikke engang ønsker der, legges til i de leksikalske ressursene som brukes i INESS.

2. Trebanker og syntaktisk analyse

Termen trebank kommer av at de første trebankene inneholdt frasestrukturtrær, som i figur 1. En slik trebank kalles gjerne en

1. INESS (Infrastructure for the Exploration of Syntax and Semantics) er et prosjekt innenfor NFRs INFRASTRUKTUR-program (Rosén et al. 2012). INESS inngår i den norske delen av CLARIN-samarbeidet, CLARINO. Se: clarino.uib.no/iness.
2. I datalingvistikken brukes termen *leksikon* (engelsk *lexicon*) om en elektronisk ordbok.

konstituenstrebank (etter engelsk *constituency treebank*), siden trærne representerer setningenes analyse i syntaktiske konstituenter.

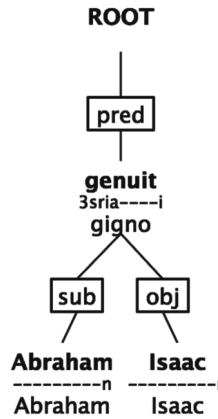


Figur 1: Syntaktisk tre for setningen *Du er meg* fra den norske Sofie-trebanken (Losnegaard et al. 2013).

Også andre typer syntaktiske strukturer forekommer i trebanker. Mange trebanker er dependenstrebanker, basert på dependens-analysen opprinnelig foreslått av Tesnière (1959). Det som står sentralt i en dependensanalyse, er avhengigheter mellom enkeltord i setningen. Et eksempel på en dependensanalyse er gitt i figur 2.

Noen trebanker er basert på bestemte lingvistiske formalismer, som for eksempel Head-driven Phrase Structure Grammar (HPSG; Pollard & Sag 1994) og Lexical Functional Grammar (LFG; Bresnan 2001, Dalrymple 2001).

Trebanker kan konstrueres på ulike måter. Det er mulig å lage en trebank helt manuelt, det vil si at annotatorer tildeler



Figur 2: Dependensstruktur for den latinske setningen *Abraham genuit Isaac* 'Abraham ble far til Isak' fra PROIEL-trebanken (Haug & Jøhndal 2008).

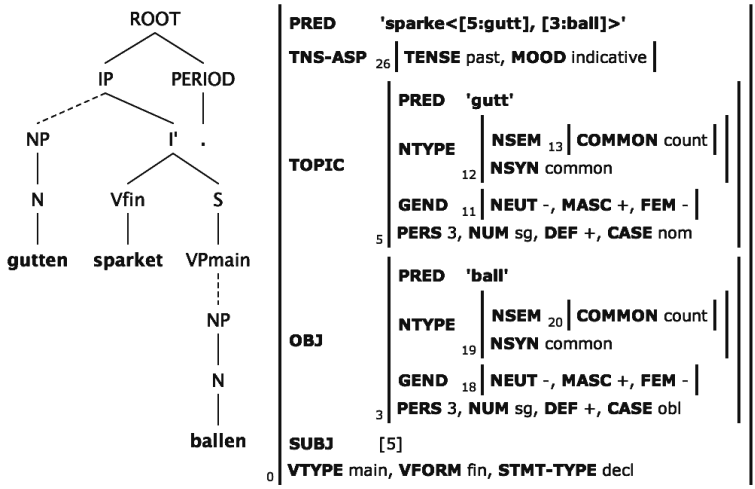
hver setning i et korpus en syntaktisk struktur. En ulempe med denne metoden er at den er svært arbeidskrevende. Det vanlige er derfor at korpuset parses automatisk. På grunn av leksikalsk og syntaktisk flertydighet vil denne metoden vanligvis resultere i flere mulige analyser for hver setning; av og til kan det være tusenvis av analyser for én enkelt setning. Derfor krever metoden at man entydiggjør, slik at man kan velge én analyse blant de foreslåtte. Entydiggjøring kan enten gjøres av annotatorer eller ved hjelp av statistiske metoder.

En viktig oppgave i INESS-prosjektet er utviklingen av en trebank for norsk. Den håndskrevne grammatikken som brukes til å parse korpuset, er NorGram (Dyvik 2000, Butt et al. 2002).[3]

3. Parsingplattform er Xerox Linguistic Environment (XLE), utviklet ved PARC (Palo Alto Research Center) i California (Maxwell & Kaplan 1993).

Strukturene i trebanken er basert på LFG, og det innebærer at hver setning får både en konstituentstruktur (c-struktur) og en funksjonell struktur (f-struktur). C-strukturen viser hvordan ordene i setningen er hierarkisk gruppert sammen i fraser. F-strukturen viser hva slags syntaktiske funksjoner frasene i setningen har, for eksempel subjekt og objekt, og også grammatiske trekk, som tall og tempus. I figur 3 vises c- og f-strukturene for setningen *Gutten sparket ballen.*

LFGs c- og f-strukturer inneholder omfattende grammatisk informasjon. Én av fordelene med trebanker, i motsetning til korpora som mangler syntaktisk annotasjon, er at det er mulig å søke ikke bare på bestemte ord, ordklasser og morfosyntaktiske trekk, men også på bestemte grammatiske konstruksjoner som ord inngår i. For eksempel kan man lett finne hvilke verb som brukes i passiv. I et tagget korpus må man søke på perfektum partisipp og eventuelt også på et hjelpe-

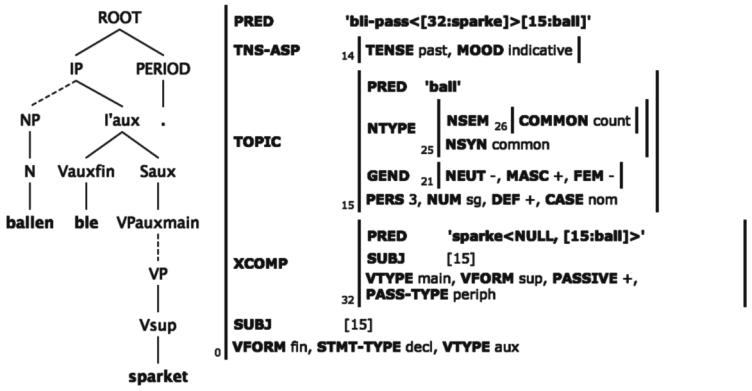


Figur 3: C- og f-struktur for setningen *Gutten sparket ballen*.

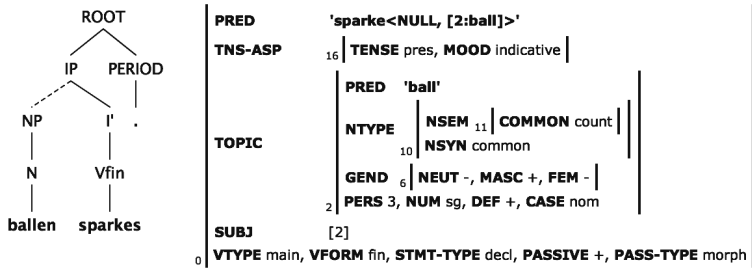
verb, eller på s-formen av verbet. Slike søk vil imidlertid gi mange treff som ikke gjelder passiv, siden både perfektum partisipp, hjelpeverbene og s-form av verb har andre funksjoner enn å danne passiv.

I en trebank kan man søke på passivkonstruksjoner på en mer direkte måte. Figur 4 viser en analyse av perifrastisk passiv, og figur 5 en analyse av morfologisk passiv. Begge analysene har verdien + for det grammatiske trekket *PASSIV* i f-strukturen. Den perifrastiske passiven har også trekket *PASS-TYPE* (passivtype) med verdi *periph* (perifrastisk), mens den morfologiske passiven har verdien *morph* (morfologisk) for dette trekket. Disse trekkene gjør det mulig å finne passiv på en enkel og treffsikker måte. For leksikografisk beskrivelse kan slike søk bidra til å klargjøre om et verb vanligvis brukes i perifrastisk passiv, morfologisk passiv eller begge deler.

Når man lager en trebank ved hjelp av en komputasjonell grammatikk, trenger man gode leksikalske ressurser. NorGram bruker flere ulike leksika i forbindelse med den syntaktiske



Figur 4: C- og f-struktur for setningen *Ballen ble sparket*.



Figur 5: C- og f-struktur for setningen *Ballen sparkes*.

analysen. De største leksikaene er basert på NorKompLeks (NKL; Nordgård 2000), et elektronisk leksikon som i sin tur er basert på *Bokmålsordboka* (*BOB:* Wangensteen 2005) og *Nynorskordboka* (*NOB:* Hovdenak et al. 2006). I tillegg til ordklasse og bøyning samt oppdelingen i lemmaer, som kommer fra *BOB* og *NOB*, inneholder NKL også informasjon om valens, altså informasjon om hvilke argumenter verb tar. Nettopp valensinformasjon er svært viktig for dyp syntaktisk analyse.

I tillegg til leksikaene som er basert på NKL, har NorGram et håndskrevet leksikon. Dette brukes bl.a. for å kode egenskapene til grammatiske ord. For at parsing skal fungere, må de syntaktiske reglene og leksikonet være samstemte med hensyn til morfosyntaktiske trekk. Som regel må de grammatiske ordene ha mer detaljert informasjon enn det som finnes i *BOB* og *NOB*.

Den norske trebanken som bygges i INESS-prosjektet, er en dynamisk trebank. Det første trinnet i arbeidet med å lage en slik trebank er å parse korpustekster automatisk. Blant de analysene som parseren foreslår, må annotatorene velge den riktige. Entydiggjøringsmetoden som brukes i INESS-trebanken, er at annotatorene velger mellom analyser gjennom å bruke diskrimi-

nanter, altså enkle egenskaper slik som entall vs. flertall, eller substantiv vs. verb. Når den ønskede analysen ikke finnes, må grammatikken og/eller leksikonet oppdateres. Etter slike oppdateringer vil en ny runde med parsing resultere i en bedre syntaktisk beskrivelse av korpuset. Denne metoden gir en dynamisk trebank der analysene kan videreutvikles parallelt med grammatikk og leksikon.

Selv med et stort leksikon og en velutviklet grammatikk får man ikke alltid en god analyse. En undersøkelse av de 255 første setningene i *Sofies verden* ble utført for å finne ut hvorfor noen setninger ikke fikk rett analyse (Losnegaard et al. 2012). Resultatet viste at 29 % av tilfellene skyldtes syntaktiske problemer, mens 71 % var leksikalske. Og blant de leksikalske problemene var de to vanligste typene flerordsuttrykk (41 % av tilfellene) og leksikalske kategorier (31 %). Resten av denne artikkelen handler om hva slags leksikalsk informasjon som er nødvendig for å få en fullverdig syntaktisk analyse.

3. Hvilken informasjon om ord er nødvendig for syntaktisk analyse?

For å få gode syntaktiske analyser trenger man korrekt informasjon om bl.a. ordklasse, valens, tellelighet, sammensetningsstruktur og flerordsuttrykk. Mye av denne informasjonen er tilgjengelig i NKL (med utgangspunkt i *BOB* og *NOB*). I INESS-prosjektet viser det seg likevel ofte at disse leksikalske

ressursene mangler informasjon som er nødvendig for å analysere ganske vanlige ord.

3.1. Ordklasse

Å hente informasjon om ordklasse fra en ordbok kan kanskje virke trivielt, men det er ikke alltid problemfritt. Det er flere grunner til at den ordklasseinformasjonen som er nødvendig, ikke alltid er tilgjengelig. Det hender at ordklassen ikke finnes i NKL fordi den var «gjemt» inne i ordartikkelen i *BOB/NOB*. Noen ganger krever syntaktisk analyse en annen ordklasse enn den som finnes i *BOB/NOB*.[4]

3.1.1. Ordklassen finnes inne i ordartikkelen

Hovedregelen for ordklassemarkering i *BOB* er at ordklassen angis rett etter oppslagsordet i artikkelhodet. I artikkelkroppen følger de ulike delbetydningene av ordet. Et eksempel gis i (1).

(1) **I abstrakt** n3 (av *II abstrakt*, i bet. 2 fra eng.) **1** substantiv med abstrakt betydning [...] **2** sammendrag, referat
II abstrakt a2 (fra lat., se *abstrahere*) **1** som bare kan fattes gjennom tanken [...] **2** språkv: som betegner en egenskap, tilstand el. handling [...]

4. Eksemplene i det følgende er tatt fra *BOB*, men forholdene er stort sett like i *NOB*.

Eksempel (1) viser at *abstrakt* har fått to oppslag markert med romertall, ett for substantivet og ett for adjektivet. Innenfor disse er ulike delbetydninger skilt fra hverandre med nummerering i arabiske tall.

I noen ordartikler står det en ny ordklassemarkering etter et arabisk tall, altså etter det som vanligvis markerer en ny delbetydning. Et eksempel gis i (2).

(2) **I absolutt** n3 (av *II absolutt*) **1** absolutt krav **2** entydig begrep **II absolutt** a2 (fra lat. av *absolvere* 'løse fra') **1** hel, fullstendig, uinnskrenket [...] **2** adv: betingelsesløst, endelig, helt [...]

For *II absolutt* er ordklassen adjektiv markert på vanlig måte rett etter oppslagsordet, men det forekommer likevel en ny ordklassemarkering for adverb etter tallet 2. Konverteringen fra *BOB* til NKL har tydeligvis ikke tatt hensyn til slike ordklassemarkeringer siden NKL bare har kategoriene substantiv og adjektiv for ordet *absolutt*.

Det er imidlertid ikke bare etter et arabisk tall at en ny ordklassemarkering kan forekomme. Eksemplene (3)–(5) viser at slik markering også kan finnes inne i definisjonsdelen av artikkelen.

(3) **jevnlig** el. **jamnlig** adv, med jevne mellomrom, stadig *besøke sine foreldre j-* / adj: hyppig *j-e besøk* [5]

(4) **juble** v1 (gj ty. fra mlat og lat., se *jubel*) rope, synge, le av glede *spillerne j-t over seieren* / *j- av glede* / adv i pr pt: *være j-nde glad*

5. I *NOB* er det omvendt: oppslaget for *jamleg* oppgir a2 som ordklasse rett etter oppslagsformen, og betegnelsen *adv* dukker opp senere i ordartikkelen.

(5) **ofte** adv (norr *opt*) hyppig, mange ganger *møtes o- / oftest* i de fleste tilfeller / *så o- (som)* konj: hver gang

I (3) er det en ny ordklassemarkering, adj, inne i ordartikkelen. I (4) er det ikke bare en ny ordklasse, adv; denne ordklassen er også knyttet til en bestemt form av oppslagsordet, presens partisipp. I (5) er det ett av eksemplene i ordartikkelen som er forsynt med en ny ordklasse, konj. I dette tilfellet ser det imidlertid ikke ut som om ordklassen gjelder selve oppslagsordet, men hele uttrykket som er oppgitt som eksempel: *så ofte (som)*. Hvis det er hele uttrykket som får en ny ordklasse, bør det kanskje få et eget oppslag (se avsnitt 3.5).

Siden ordklassemarkeringene inne i ordartiklene er gjort på så mange ulike måter, kan det være vanskelig å finne alle ved konvertering til et elektronisk leksikon. Også brukere av ordboken i papirformat kan tenkes å ha problemer med å finne frem til relevant ordklassebetegnelse når den kan være så dypt innføyet i ordartikkelen.

3.1.2. Ordklassen er ikke optimal for syntaktisk analyse

Noen ord har en ordklasse som passer dårlig for automatisk syntaktisk analyse. Et eksempel finnes i oppslaget for verbet *unnskylde*.

(6) **unnskylde** v2 (fra lty, sm o s ty. *entschuldigen*, eg 'frita for skyld') **1** beklage, be om unnskyldning for *u- sin glemsomhet* / ofte refl: *hun unnskyldte seg fordi hun kom så sent* **2** som høflig innledning til et spørsmål e l: *unnskyld, kunne De si meg veien til byen?* [...]

Her er analysen at *unnskyld* som «høflig innledning til et spørsmål eller lignende» er imperativ av verbet *unnskylde*. Dette er i og for seg en mulighet, men for automatisk syntaktisk analyse vil det være fordelaktig å også kunne analysere ordet som en interjeksjon. Grunnen er at *unnskyld* ofte har den syntaktiske distribusjonen til en interjeksjon, noe som ikke er tilfelle for imperativ av hvilket som helst verb. I denne bruksmåten er det også svært begrenset hvilke argumenter ordet kan ha. Objekt er en mulighet, slik som i *unnskyld meg* eller *unnskyld oss*, men ellers er det få mulige utfyllinger.

Noen ganger trenger man en mer finkornet inndeling enn den de tradisjonelle ordklassene gir. For eksempel er det i NKL bare en stor klasse adverb, som inneholder mange svært uensartede ord. Siden ulike typer adverb har ulik syntaktisk distribusjon, er det nødvendig å klassifisere dem i underkategorier.

(7) *Foredraget var **temmelig** langt.*
(8) *Han liker **dessverre** ikke ostekake.*
(9) *Vi har **vel** hørt den før.*

Eksemplene (7)–(9) illustrerer henholdsvis gradsadverb, holdningsadverb og partikkeladverb. Partikkeladverb kommer før holdningsadverb, slik at *Vi har vel dessverre hørt den før* er mulig, mens **Vi har dessverre vel hørt den før* ikke er det. Når annotatorene finner at en setning med et adverb får en utilfredsstillende analyse, kan dette ofte løses gjennom at ordklassen endres fra default-kategorien adverb til en av de mer finkornete adverbkategoriene.

3.2. Valens

For å få en god syntaktisk analyse av en setning er det nødvendig å ha passende valensrammer for verbene i setningen. NKL angir valensrammer for alle verb, og disse utgjør grunnstammen i valensrammene som brukes i NorGrams leksikon. Når et korpus er parset, finner annotatorene ofte at rammer mangler, og disse legges da til i NorGrams leksikon. Eksemplene i (10) og (11) er setninger som ikke fikk korrekt analyse fordi den nødvendige valensrammen manglet.

(10) *Faren **mumlet** et farvel.*
(11) *Han **trengte** seg fram.*

Verbet i (10) manglet en transitiv ramme, og verbet i (11) manglet en ramme med refleksivt objekt og partikkel.

Også substantiv og adjektiv kan ta komplementer, som vist i (12) og (13).

(12) *Orker ikke **tanken** på mer drittslenging.*
(13) *Han er **stolt** av datteren.*

En ramme der substantivet *tanke* tar en preposisjonsfrase er nødvendig for å analysere (12), og en lignende ramme for adjektivet *stolt* er nødvendig for (13).

Noen verb kan forekomme i såkalte inquit-konstruksjoner, der et sitat følges av finitt verb og subjekt (og eventuelt andre setningsledd). Forbausende mange verb forekommer i denne konstruksjonen; noen eksempler vises i (14)–(18).

(14) *– Jeg finner da veien hjem, **skrek** Fredrik etter ham.*

(15) *– Ja, jeg leter etter en jobb,* ***lyver*** *jeg.*
(16) *«Jeg vil se Farid,»* ***jamret*** *Zoubida.*
(17) *Det dufter deilig,* ***skrøt*** *han mens hun bakte julekaker.*
(18) *«Det er telegram fra doktern, fra Alex,»* ***rettet*** *hun.*

Inquit-verb er et subsett av verb som tar *at*-setninger som komplement; verb som *glemme*, *lære*, *oppnå* osv. er vanskelig å tenke seg i denne konstruksjonen. Siden NorGram-leksikonet ikke i utgangspunktet har informasjon om hvilke verb som kan brukes i inquit-konstruksjoner, må disse legges til når de dukker opp under entydiggjøringsarbeidet.

3.3. Tellelighet

Informasjon om tellelighet er relevant for syntaktisk analyse, siden tellelige og ikke-tellelige substantiv til dels har ulik syntaktisk distribusjon. For eksempel kan ikke-tellelige substantiv brukes som komplette nomenfraser uten determinativ i ubestemt entall, mens dette vanligvis ikke er mulig for tellelige.

(19) *Øl/Kjøtt/Ertepuré er i kjøleskapet.*
(20) **Drink/Kjøttkake/Ert er i kjøleskapet.*

Siden det ikke finnes informasjon om tellelighet i NKL, er alle substantiv blitt tildelt trekket tellelig som default i NorGram-leksikonet. Når ikke-tellelige substantiv oppdages under entydiggjøring, legges denne informasjonen til. Eksempler funnet i trebanken er gitt i (21) og (22).

(21) *Piken var en skjensel, hun kastet **vanry** over dem.*
(22) *Han løftet forsiktig opp et smykke som glimtet i **gull**.*

Selv om informasjon om tellelighet ikke oppgis i *BOB*, er det noen substantiv som har entallskoder, som for eksempel *brie, øst, pomp, rytmikk* osv. Det er mulig at denne morfologiske informasjonen kan utnyttes til å finne en del ikke-tellelige ord i *BOB*, men informasjonen er ikke angitt systematisk.

3.4. Sammensetningsstruktur

I norsk er sammensetning en produktiv ordlagingsmåte. Nye sammensetninger lages i stor grad spontant i tale og i skrift, og disse kan naturligvis ikke føres opp i vanlige ordbøker. Før en tekst kan parses, må den analyseres morfologisk. Produktive sammensetninger kan da håndteres ved hjelp av automatisk sammensetningsanalyse.

Leksikaliserte sammensetninger har sin naturlige plass i ordbøker. I *BOB* skiller disse seg ikke ut fra usammensatte ord; det er ingen informasjon om den indre morfologiske strukturen til leksikaliserte sammensetninger. Men det kan være fordelaktig å ha slik informasjon, bl.a. for å kunne analysere elliptiske koordinasjoner av den typen som vises i (23)–(25).

(23) *munn- og klovsyke*
(24) *Norsk Nærings- og Nytelsesmiddelarbeiderforbund*
(25) *vinter- og sommerdekk*

I (25) gjelder koordinasjonen leddene *vinter-* og *sommer-*, ikke

leddet *vinter-* og ordet *sommerdekk*. Hvis man skal kunne få en tilfredsstillende syntaktisk og semantisk analyse av disse, må man vite at *sommerdekk* består av leddene *sommer* og *dekk*.

3.5. Flerordsuttrykk

Flerordsuttrykk utgjør en stor utfordring for automatisk syntaktisk analyse. Hvis de behandles som sekvenser av enkle ord, får man ofte uriktige analyser. Det store problemet er at det er få kilder til kunnskap om hvilke flerordsuttrykk som finnes. Jackendoff anslo at antallet flerordsuttrykk i engelsk er av omtrent samme størrelsesorden som antallet enkle ord (1997: 156). Selv om det er vanskelig å anslå dette for norsk, er det liten tvil om at det også i norsk er et meget høyt antall flerordsuttrykk.

Det finnes flere forskjellige typer flerordsuttrykk; Sag et al. gir en god oversikt (2002). Én av de viktigste typene er faste uttrykk som er leksikalisert, og som ikke kan bøyes eller endres på andre måter. Disse kalles gjerne ord med mellomrom (*words with spaces* på engelsk), og det er lett å behandle dem som egne oppslag. *BOB* har en del oppslag av denne typen, for eksempel: *ad absurdum, for så vidt, i hende, lille julaften, tipp topp, world cup* osv. Det er nok mange flere slike uttrykk som ikke har egne oppslag, men som ordboken har informasjon om. Disse omtales inne i ordartiklene til ett eller flere av de grafiske ordene som inngår i dem. Et eksempel er uttrykket *dann og vann*, se (26)–(28).

(26) **I vann** el. **vatn** n1 (norr *vatn*) **1** klar, gjennomsiktig væske [...]

(27) **II vann** adv, se *dann*
(28) **dann** adv (ty. *dann und wann*) bare i uttr *d- og vann* nå og da

Ordet *vann* har altså to oppslag, ett som substantiv og ett som adverb. Oppslaget som adverb henviser til oppslaget for *dann*, som også er kategorisert som adverb. Men hverken *vann* eller *dann* kan brukes som adverb alene i norsk. Det faste uttrykket *dann og vann* fungerer som adverb, og det fortjener et eget oppslag. For automatisk syntaktisk analyse er det uheldig at *vann* er kategorisert som et adverb; dette betyr at når ordet forekommer i en setning, vil parseren ofte foreslå en helt uaktuell mulighet, nemlig at ordet *vann* fungerer som adverb. Heller ikke for menneskelige brukere av en papirordbok kan det ha noe for seg at enkeltordene *vann* og *dann* har egne oppslag som adverb.

Under oppslaget for *god* finnes eksempler som skal illustrere substantivisk bruk.

(29) **god** a1 [...] **1** av høy kvalitet, bra, fin, gagnlig, tjenlig, skikket, dugende *g-t vær* / *g-e veier, forhold* / *en g- film, bok, debatt* / *ha g- helse, hørsel, samvittighet* / *et g-t spørsmål* / *ha g-e intensjoner* / *en g- kniv* / *i g-e, gamle dager* [...] / subst, med gl gen. etter til: *tvilen kommer tiltalte til g-e* blir godskrevet tiltalte / *(ha) til g-e* til overs, utestående; (ha) til senere / *g-t!* fint! / subst: *hva sier han til g-t?* hva har han å si? [...]

De formene som skal være substantiv, er altså *gode* i uttrykket *til gode* og *godt* i uttrykket *(hva sier han) til godt*. For at det skal være berettiget å kategorisere disse formene som substantiv, bør de kunne fungere i syntaktiske posisjoner som er typiske for substantiv. Disse eksemplene viser at de riktignok

kan forekomme etter en preposisjon, og dette er en typisk syntaktisk posisjon for nomenfraser. Men disse formene forekommer knapt etter andre preposisjoner enn *til*, og de kan ikke bygges ut med andre ord som er typiske for nomenfraser. Det virker mer nærliggende å betrakte dem som faste uttrykk som bør ha egne oppslag.

Noen flerordsuttrykk blir oppdaget under entydiggjøring fordi annotatorene merker at det er noe som ikke stemmer med analysen. Et eksempel på en slik setning gjengis i (30).

(30) *Men før Artur hoppet **over bord**, hadde Martin hatt et hav av tid.*

Problemet her er at det er noe rart med preposisjonsfrasen *over bord*, siden objektet for preposisjonen er et nakent substantiv. Man kan si *han hoppet over stolen/over relingen/over bordet*, men ikke *han hoppet over stol/over reling*. Når det går bra med *over bord*, er det fordi det er et stivnet uttrykk med en leksikalisert betydning. Uttrykket er med som eksempel i *BOB*, men det har ikke et eget oppslag.

Av og til oppdages flerordsuttrykk under entydiggjøring fordi et uttrykk som etter gjeldende normering skal skrives som flere grafiske ord, er blitt sammenskrevet. Et eksempel er *værsågod*. Flerordsuttrykket *vær så god* er markert som uttrykk i *BOB* under oppslagene for alle ord som inngår i det, både *god*, *så* og *være*.

Det kan være interessant å sammenligne med behandlingen dette uttrykket får i noen danske og svenske ordbøker. Oppslaget i (31) er fra *Politikens Nudansk Ordbog* (Becker-Christensen 1999).

(31) **værsgo** [...] udråbsord **1**. udtryk som bruges når man giver

> el. rækker andre noget □ *værsgo, maden er serveret! · værsgo, her er pengene · vil du række mig saltet? - ja selvfølgelig, værsgo!*

Her er uttrykket trukket sammen til ett ord. I alle eksemplene i denne ordartikkelen passer det fint med analysen som interjeksjon, eller «udråbsord», som det heter her. Det typiske for interjeksjoner er at de står for seg selv, egentlig utenfor integrerte syntaktiske konstruksjoner. Her er alle adskilt fra resten av setningen med komma, enten i begynnelsen eller slutten av setningen.

Oppslaget i (32) er fra *Dansk-dansk Ordbog* (Dissing & Helles 1992).

(32) **værsgo** (udråbsord) (el. *vær så god*). Værsgo maden er serveret! (forstærkende:) Vil du værsgo (el. vær så god) gøre rent efter dig.

I det første eksempelet i dette oppslaget har ordet *værsgo* en posisjon som er typisk for interjeksjoner, men i det andre, skrevet enten som ett ord eller som flere, er plasseringen heller typisk for adverb.

Oppslaget i (33) er fra *Norstedts stora svensk-engelska ordbok* (Sjödin 2000).

(33) **varsågod** se under *god I 1*

Flerordsuttrykket *var så god* behandles utførlig i artikkelen under *god*, men med en henvisning fra et enkeltordsoppslag. Et slikt oppslag kan være spesielt viktig i en tospråklig ordbok, der brukere kan tenkes å ha vanskelig for å finne frem til ordet hvis det ikke har sitt eget oppslag.

Det ser ut til at dette ordet/uttrykket kan forekomme både

som enkeltord og som flerordsuttrykk i alle de skandinaviske språkene. I *BOB* er det kanskje slik at et enkeltordsoppslag mangler fordi normering har tilsagt at uttrykket skal skrives med flere ord. På den annen side er det god grunn til å behandle uttrykket komposisjonelt i visse tilfeller; ett av eksemplene under oppslaget for *god* er: *vær så g- å gå til bords.* Men tatt i betraktning at dette er et så høyfrekvent ord i dagligspråket, ofte uttalt med stor grad av fonologisk reduksjon, er kriteriene absolutt til stede for at det skal ha et oppslag som enkeltord. Dette virker som et tilfelle der det er høyst berettiget å ha begge typer oppslag.

4. Hvilke ord er nødvendige for syntaktisk analyse?

Når man parser et korpus, vil det alltid være ord som er ukjente for den morfologiske komponenten og/eller for leksikon. Det kan være tale om nyord, men det kan også dreie seg om feilstavede ord, feil bøyningsmåte osv. Ett eneste ukjent ord i en setning vil kunne resultere i at parseren ikke finner rett analyse. For å få til en vellykket syntaktisk analyse må flest mulig av de ordene som forekommer i autentiske tekster, gjenkjennes, enten de er korrekte eller ikke.

INESS har et eget grensesnitt for tekstpreprosessering. Dette grensesnittet har to viktige funksjoner. Det første er å oppdage og korrigere feil som har oppstått på grunn av optisk tegngjenkjenning (OCR) i skannete dokumenter som leveres fra Nasjonalbiblioteket. Det andre er å registrere og klassifisere ukjente ord i de samme dokumentene.

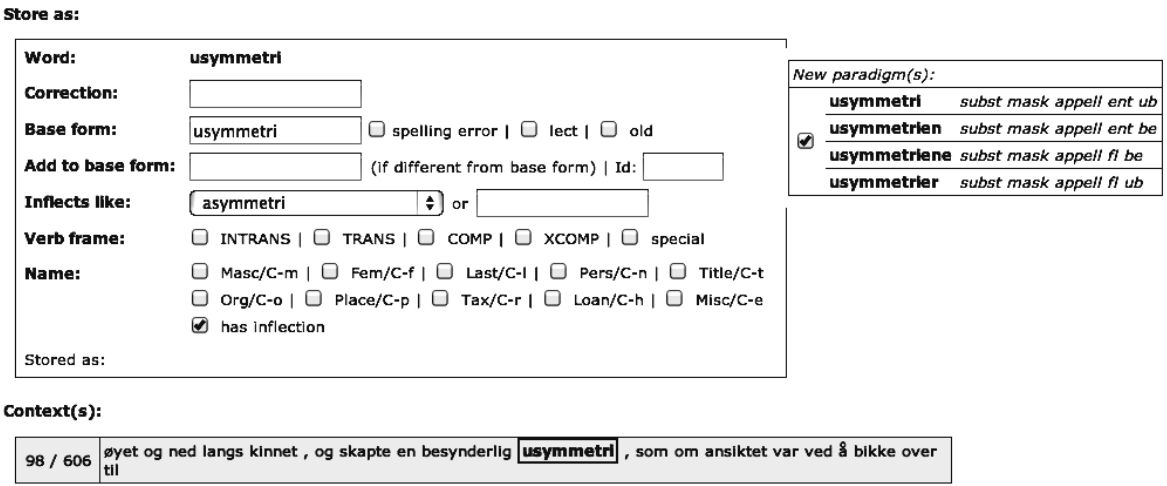
Preprosesseringen av tekstene finner alle grafiske ord som ikke kjennes igjen av den morfologiske komponenten. Disse ordene blir så presentert for annotatorene. Hvis ordet er en OCR-feil, korrigeres den av annotatoren. For eksempel skjer det ganske ofte at ordet *lo* tolkes som tallet *10*. Hvis ordet er korrekt skannet, men likevel ikke gjenkjennes av den morfologiske analysen, behandles det på ulike måter avhengig av hva slags problem det gjelder.

Noen ganger er ordet en neologisme som ikke finnes i leksikon. Figur 6 viser grensesnittet som brukes for å legge til et nytt ord. I dette eksempelet er det ordet *usymmetri* som skal legges inn. Annotatoren legger til ordet gjennom å oppgi *usymmetri* som baseform. Systemet foreslår da et antall ord som ligner på denne strengen, nemlig de ordene det allerede har som ender på samme måte. Fra en rullegardinmeny velger annotatoren et ord som bøyes på samme måte som det nye ordet. Så kommer det opp et nytt vindu (til høyre i bildet) med et komplett bøyningsparadigme, slik at annotatoren kan sjekke at paradigmet er rett.

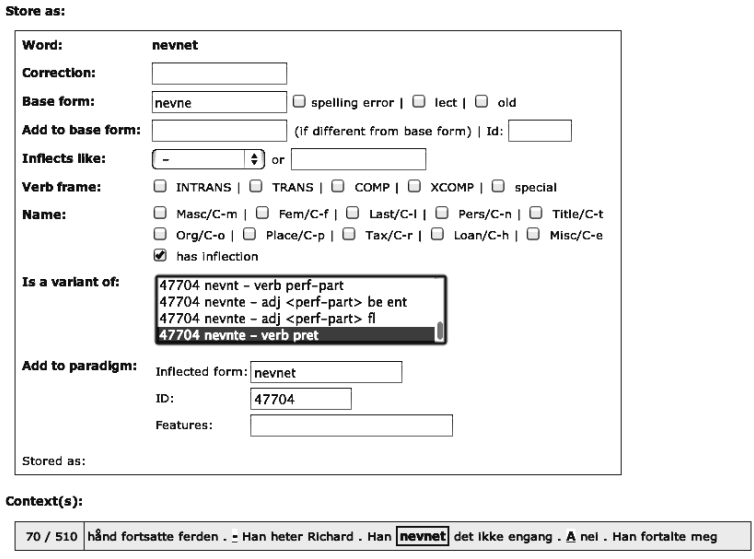
Noen ganger er ordet kjent, men bøyningsmåten er uvanlig. Et slikt eksempel er *nevnet* som preteritum av verbet *nevne*. Figur 7 viser grensesnittet under tillegg av en bøyningsform. Baseformen spesifiseres, og systemet viser bøyningsparadigmet. Annotatoren kan da velge hvilken av bøyningsformene i det vanlige paradigmet som den nye formen er en variant av. Denne formen legges så til bøyningsparadigmet for ordet.

Det er ofte feilstavede ord i tekstene som analyseres. For at også setninger med slike ord skal kunne parses, kan disse legges inn som varianter. I figur 8 er det et eksempel på det feilstavede ordet *alldeles* som legges inn.

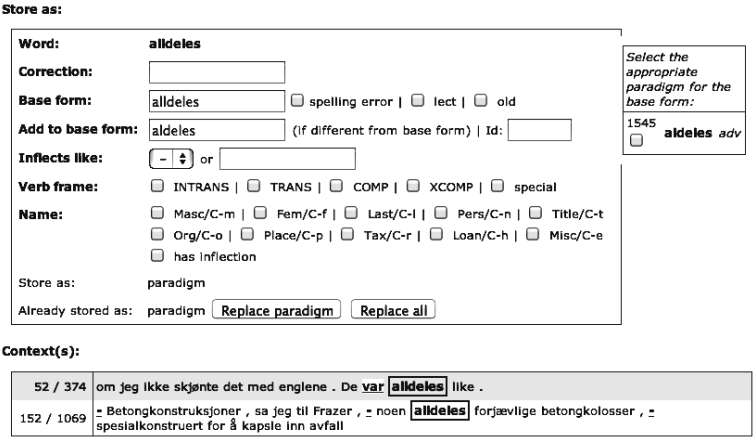
En ordklasse med mange kreative nydannelser er interjeksjonene, både sammensetninger, som i (34), og infikser, som i (35).



Figur 6: Skjermbilde av grensesnittet for et ukjent substantiv.



Figur 7: Skjermbilde av grensesnittet for valg av ny bøyningsform.



Figur 8: Skjermbilde av grensesnittet under innlegging av alternativ stammeform.

(34) *Kan få fliser i tunga! Og kvae!* ***Dobbelt-æsj****! Verste som finnes!*
(35) *[…] og hanen Hanibal gol: «****Kykkeli-gratuly-ky****»!*

Ofte er ordet i og for seg kjent, men skrivemåten er annerledes enn den normerte. I tabell 1 er det eksempler på interjeksjoner med stavemåten(e) som forekommer i korpuset, sammenstilt med det som ser ut til å være normert skrivemåte. Det ganske vanlige ordet *jippi* står hverken i *BOB* eller *NOB*, men Kunnskapsforlagets *Norsk-engelsk stor ordbok* har det som oppslag (Henriksen & Haslerud 2002). Særlig skrivemåten *fy til rakkeren* skiller seg ut; dette virker mer som en etymologi enn en avspeiling av hvordan dette muntlige uttrykket uttales.

Formene funnet i korpuset	Normerte former
jovisst	*jo visst*
jaja	*ja ja*
jøssda	*jøss da*
javelja	*ja vel ja*
uffda	*uff da*
heisann/heisan	*hei sann*
ojsann/oisan	*oi sann*
fyttirakkern/fytterakker'n/fytte rakker'n	*fy til rakkeren*
jipppiii/jippiiiii/jipppiiii	*jippi*

Tabell 1: Skrivevarianter av noen interjeksjoner.

Norske forfattere bruker ofte ord fra andre språk i tekstene sine. Engelske ord er spesielt populære, som i talespråket, men

også ord fra andre språk brukes flittig, for eksempel i skjønnlitterære dialoger. Noen eksempler som er blitt registrert under tekstpreprosessering er gjengitt i (36)–(40).

(36) *Er han en slags, he-he,* ***boy-friend*** *eller noe sånt? spurte han.*
(37) *Jeg dro rundt med mitt kamera og min «****business class****»-billett på maven.*
(38) «***Au contraire****, Nick,» lød Edmonds lyse, slepende stemme fra døren.*
(39) *– Jeg har med frisisk* ***kruidkoek*** *til deg, sa Natasha og la en krydderkake på bordet.*
(40) *Men dette greier du jo fint, det er jo* ***peanuts*** *for deg, Halvdan, […]*

For at slike ord skal få plass i en vanlig ordbok, vil man kreve at de er godt etablerte som lånord i språket. Men for automatisk analyse er det avgjørende at alle ord gjenkjennes. Derfor legges også slike ord inn i morfologi og leksikon. Det siste eksempelet, *peanuts*, er så vanlig i norsk at det kan synes merkelig at det ikke har fått innpass i ordbøkene.

5. Konklusjon

For at en tekst skal kunne analyseres automatisk, må alle ord i teksten, enten de er en del av standardvokabularet eller de er lånord, feilstavinger, unormerte bøyningsformer e.l., være kjent

for systemet. I en vanlig ordbok pleier man å kreve et visst antall forekomster som et kriterium for et nytt leksikalsk oppslag, men for formålet parsing av autentisk tekst kan én forekomst være nok til å rettferdiggjøre et nytt oppslag. I denne artikkelen er det brukt eksempler som er relevante for bygging av en trebank, men dette er langt fra den eneste applikasjonen som har bruk for ord som ikke får plass i konvensjonelle ordbøker. For eksempel kan en liste over vanlige feilstavinger være nyttig for stavekontrollprogrammer.

I forbindelse med utviklingen av den norske trebanken i INESS utfører prosjektet et omfattende arbeid med leksikalske ressurser. Resultatene av dette arbeidet kan være av interesse for leksikografer. Fremtidsvisjonen bør være at vi kan utvikle et tettere samarbeid mellom de leksikografiske og språkteknologiske miljøene, til gjensidig glede og nytte.

Litteratur

Becker-Christensen, Christian (1999): *Politikens Nudansk Ordbog*. Aalborg: Politikens Forlag.

Bresnan, Joan (2001): *Lexical-Functional Syntax*. Oxford: Blackwell.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi & Christian Rohrer (2002): The Parallel Grammar project. I: John Carroll, Nelleke Oostdijk & Richard Sutcliffe (red.): *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, *Taipei, Taiwan*.

Stroudsburg, Pennsylvania: Association for Computational Linguistics.

Dalrymple, Mary (2001): *Lexical Functional Grammar*, volume 34 of Syntax and Semantics. San Diego, California: Academic Press.

Dissing, Børge & Sigrid Helles (1992): *Dansk-dansk Ordbog.* København: Gyldendal.

Dyvik, Helge (2000): Nødvendige noder i norsk. Grunntrekk i en leksikalsk-funksjonell beskrivelse av norsk syntaks. I: Øivin Andersen, Kjersti Fløttum & Torodd Kinn (red.): *Menneske, språk og felleskap*. Oslo: Novus forlag.

Haug, Dag T. T. & Marius L. Jøhndal (2008): Creating a parallel treebank of the old Indo-European Bible translations. I: *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008), Marrakech, Morocco, 1st June 2008*.

Henriksen, Petter & Vibecke C.D. Haslerud (2002): *Norsk-engelsk stor ordbok*. Oslo: Kunnskapsforlaget.

Hovdenak, Marit, Laurits Killingbergtrø, Arne Lauvhjell, Sigurd Nordlie, Magne Rommetveit & Dagfinn Worren (2006): *Nynorskordboka: Definisjons- og rettskrivingsordbok*. Oslo: Det Norske Samlaget.

Jackendoff, Ray (1997): *The architecture of the language faculty*. Cambridge, Mass.: MIT Press.

Losnegaard, Gyri Smørdal, Gunn Inger Lyse, Anje Müller Gjesdal, Koenraad De Smedt, Paul Meurer, & Victoria Rosén (2013): Linking Northern European infrastructures for improving the accessibility and documentation of complex resources. I: Koenraad De Smedt, Lars Borin, Krister Lindén, Bente Maegaard, Eiríkur Rögnvaldsson, & Kadri Vider, (red.): *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013*.

Linköping: Linköping University Electronic Press. <http://www.ep.liu.se/ecp/089/005/ecp1389005.pdf> (mars 2014).

Losnegaard, Gyri Smørdal, Gunn Inger Lyse, Martha Thunes, Victoria Rosén, Koenraad De Smedt, Helge Dyvik & Paul Meurer (2012): What we have learned from Sofie: Extending lexical and grammatical coverage in an LFG parsebank. I: Jan Hajič, Koenraad De Smedt, Marko Tadič & Antonio Branco, (red.): *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*. European Language Resources Association, 69–76. <http://link.uib.no/?rosen12lrec> (mars 2014).

Maxwell, John & Ronald M. Kaplan (1993): The interface between phrasal and functional constraints. I: *Computational Linguistics*, *19(4)*:571–589.

Nordgård, Torbjørn (2000): NorKompLeks – A Norwegian computational lexicon. I: *COMLEX 2000*, Patras, Greece.

Pollard, Carl & Ivan A. Sag (1994): *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Rosén, Victoria, Koenraad De Smedt, Paul Meurer & Helge Dyvik (2012): An open infrastructure for advanced treebanking. I: Jan Hajič, Koenraad De Smedt, Marko Tadić, & António Branco (red.): *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*. European Language Resources Association, 22–29. <http://link.uib.no/?rosen12lrec> (mars 2014).

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger (2002): Multiword expressions: A pain in the neck for NLP. I: Alexander Gelbukh (red.): *Computational linguistics and intelligent text processing: third international conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002: proceedings.* Berlin: Springer, 189–206.

Sjödin, Maria (2000): *Norstedts stora svensk-engelska ordbok*. Stockholm: Norstedt.
Tesnière, Lucien (1959): *Éléments de syntaxe structurale*. Paris: Klincksieck.
Wangensteen, Boye (2005): *Bokmålsordboka: Definisjons- og rettskrivningsordbok*. Oslo: Kunnskapsforlaget.

Victoria Rosén
førsteamanuensis, dr.art.
Universitetet i Bergen
Sydnesplassen 7
N-5007 Bergen
victoria@uib.no