Knut Hofland
NAVFs edb-senter for humanistisk forskning
Postboks 53
N-5014 Bergen-Universitet

GRAMMATISK MERKING AV LOB-KORPUS.

Innledning

Denne artikkelen omtaler det pågående arbeid med grammatisk merking (på ordklassenivå) av LOB (Lancaster-Oslo/Bergen) korpus. Dette er et samarbeidsprosjekt mellom Britisk Institutt i Oslo, NAVFs edb-senter i Bergen og universitetet i Lancaster. På grunn av reduserte bevilgninger vil hovedtyngden av arbeidet bli gjort i Lancaster.

LOB korpus er en britisk-engelsk parallell til det amerikanske Brown korpus som ble ferdig i 1967. For opplysninger om Brown korpus se Francis (1979) og Kucera & Francis (1967). Arbeidet med LOB korpuset ble startet i Lancaster i 1970 av Geoffrey N. Leech og ble fullført i Norge ved et samarbeid mellom Stig Johansson, Oslo og NAVFs edb-senter, Johansson et al. (1978). Til LOB korpus er det laget en KWIC-konkordans på mikrokort, Hofland & Johansson (1979) og det er under utgivelse bearbeidete ordlister til materialet bl. a. med sammenligninger med Brown korpus, Hofland & Johansson (1981).

Brown og LOB korpora inneholder hver 500 tekstutsnitt på omlag 2000 ord, totalt 1 million ord, fra forskjellige typer trykket tekst (inndelt i 15 kategorier) utgitt i 1961. De to korpora har samme oppbygging og egner seg derfor godt til sammenligninger. I 1979 ble det gjort ferdig en grammatisk merket versjon av Brown korpus. En analyse av dette materialet av Kucera & Francis er under utgivelse.

Arbeidet med å finansiere et tilsvarende prosjekt for merking av LOB korpuset ble startet i 1979. Forarbeidet ble påbegynt i 1980 og fra 1981 er det bevilget midler, i England også for 1982. Merkingen av teksten gjøres av følgende grunner:

a) den gjør det mulig å søke i tekst både etter kombinasjon av ord og grammatiske koder
b) separerer homografer
c) gjør senere lemmatisering av materialet enklere
d) gjør mulig automatisert syntaktisk analyse av materialet

Metode

For også å kunne sammenligne de merkete korpora vil det samme

Grammatisk merking av LOB-korpus
Knut Hofland
Proceedings of NODALIDA 1981, pages 117-124

kodesystemet som ble brukt ved merkingen av Brown korpus bli brukt ved merkingen av LOB korpus. Merkesystemet gir i hovedsak koder basert på ordklassetilhørighet og kodene kan deles inn i 5 typer (se fullstendig kodeliste i appendix a):

a) åpne ordklasser
b) funksjonsord (lukkete ordklasser)
c) viktige enkle ord som not, be, have, do
d) tegnsetting som kan ha syntaktisk informasjon
e) bøyingsmorfemer til a) og c)

Metoden som skal brukes er en modifisert utgave av den som ble brukt ved merking av Brown korpus, Greene & Rubin (1971). Denne består av 5 trinn:

1) pre-editering av teksten
2) oppslag i ordliste (Brown: 2860 ord, 61% med en kode)
3) oppslag i endelsesliste basert på inntil 5 siste tegn, (Brown: 446 endelser, 51% med en kode)
4) kontekstregler innen setning, brukt på ord som har fått flere koder etter trinn 2 og 3 (Brown: 77% rette koder)
5) Manuell entydiggjøring og oppretting av feile koder

Arbeidet med å tilpasse endelseslisten til britisk-engelsk og utvidelse av denne og ordlisten er beskrevet av Mette-Cathrine Jahr i neste artikkel.

Ved modifikasjon av metode og tabeller har følgende materiale vært tilgjengelig:

1) Det merkete Brown korpus, og følgende lister
   a) alfabetisk ordliste med tilhørende koder
   b) finalalfabetisk ordliste med tilhørende koder
   c) alfabetisk liste over koder med tilhørende ord
   d) maskinell produsert endelsesliste basert på endelser med koder som forekommer i minst 5 ord
2) LOB korpus med ordlister og KWIC-konkordans
3) Finalalfabetisk ordliste til både Brown og LOB korpus (totalt ca. 75 000 grafiske ord)
4) Liste av ord som forekommer i begge korpora (ca. 25 000)
5) Liste av ord som bare forekommer i LOB korpus (ca. 25 000)

I forbindelse med forberedelsen til prosjektet er en del vanlige homografer som may, to, that, merket manuelt utifra den eksisterende KWIC-konkordans. Merkene overføres til teksten og vil særlig få betydning ved bruk av kontekstreglene.

Pre-editering av teksten.

Ved kodingen av Brown korpus ble teksten redigert ved at en del tegnsetting og koder ble fjernet. Noen ord ble slått sammen til enheter som f.eks. navn på personer og organisasjoner, datoer ol. Disse fikk spesielle koder tilordnet. Stor forbokstav ble bare beholdt for egennavn. I LOB korpuset kan en del av de eksisteren-

de koder brukes i dette arbeidet. Det gjelder markering av setningstart, koder for forkorting og utenlandske ord, overskrifter ol.

Endelseslisten

Denne listen inneholder tradisjonelle avledningss- og bøyingssuffikser, men også endelser som kan identifisere en ordklasse uten at den har noen grammatisk funksjon. I endelseslisten ønsker en "de (lengste) endelser som identifiserer færrest mulig ordklasser (helst bare en) og så mange ord som mulig". Eksempel fra listen:

```
   IVE      --> JJ - NN    (adjektiv eller substantiv)
 CEIVE      --> VB         (verb)
  RIVE      --> VB
  SIVE      --> JJ
  TIVE      --> NN - JJ
  VIVE      --> VB
```

Den lengste endelse som fins i listen brukes.

Ordlisten

Ordlisten inneholder funksjonsord og ord som ikke følger endelseslisten eller de spesialbehandlinger som foretas. Videre inneholder den alle ord med frekvens 50 eller mer i Brown korpus. Ved merking av LOB korpuset er hele ordlisten fra det merkete Brown korpus tilgjengelig. Men noen av ordene der som bare har fått en kode kan allikevel være homografer slik at ord fra denne ordlisten må spesialbehandles. En mulighet er i tillegg å slå opp i endelseslisten for å finne eventuelle andre koder som kan forekomme.

Spesialbehandling av en del ord

Programmet til Greene & Rubin foretar først oppslag i ordliste. Hvis ordet forekommer der velges koden(e) fra ordlisten, ellers foretas en sjekk etter spesielle ord.

a) ord som begynner med $ merkes NNS.
b) For ord som inneholder apostrof fjernes endingen (N'T, 'LL, 'RE, 'VE, 'D, 'S, ') og resten av ordet sjekkes. Koder for endingen henges på som en tilleggskode.
c) koder som er satt på under pre-editering overføres til de enkelte ord
d) ord med stor førstebokstav får kode NP.
e) ord med bindestrek splittes opp i to deler. Som regel får ordet koden til siste ledd. En del kombinasjoner av koder og

endelser for de to delene behandles spesielt.
f) Ord som starter med et siffer får kode CD. Tall skrevet med bokstaver må stå i ordliste unntatt -TEEN.
g) Ord som slutter på ST, RD, ND og som har siffer som første tegn får kode OD.
h) Ord på formen UN...ED merkes JJ. Der hvor formen UN... er et verb må dette stå i ordlisten.

Ord som ikke slutter på enkel S sjekkes mot endelseslisten. Hvis endelsen ikke fins der, får ordet kode NN - VB - JJ.

Ord som slutter på S får spesialbehandling. Ordet gies foreløpig kode NNS eller VBZ. S'en fjernes fra ordet og dette sjekkes i ordlisten. Hvis det fins der velges en av mulighetene, NNS hvis NN og ikke VB fra ordlisten, VBZ hvis VB og ikke NN fra ordlisten.

Dersom ikke ordet uten S fins i ordlisten sjekkes endelsen til ordet.

a) ved -ING gis kode NNS
b) ord som ender på konsonant sjekkes i endelseslisten. Ord som slutter på konsonant+S må stå i ordlisten
c) -IE forandres til Y og ordet sjekkes i ordliste
d) ved ord som slutter på -SES, ZES, HES, XES sjekkes ordet uten -ES i ordlisten
e) ord som slutter på I(S) gis kode NN, EAU(S) kode NNS, OU(S) kode JJ, U(S) kode NN
f) ellers sjekkes mot ordliste/endelsesliste

I tillegg til denne spesialbehandlingen av S kan det være aktuellt også å fjerne endelser som -ISH, -ED, -(E)R, -LY og sjekke resten av ordet mot ordliste og endelsesliste. Dersom f.eks. et ord slutter på -ISH og resten av ordet ikke er et verb, så gis ordet kode JJ.

Kontekstregler

Når alle ordene i en setning har fått koder, skal kontekstreglene velge ut riktig kode der et ord har fått flere koder, basert på kodene til de omsluttende ord (for Brown korpus inntil 2 ord på hver side). For at reglene skal kunne brukes, må et eller flere av de omsluttende ord ha en entydig kode. Kontekstreglene kan være av 2 typer:

1) negative, f.eks. at VB ikke kan etterfølge AT

   AT ? --> -VB

2) positive, f.eks etter en modal kan det komme verb i grunnform

   MD ? --> VB

Ved utarbeidingen av kontekstreglene til Brown korpuset ble 900 setninger merket manuelt. Det ble kjørt ut sorterte lister over kombinasjoner av koder og kontekstreglene ble satt opp etter dette grunnlaget. Kontekstreglene til LOB korpuset vil bli laget på grunnlag av hele det merkete Brown korpuset.

Totalt kan det være 8 mulige regler for et ord med flere koder

```
1)  A  B  ?  C  D
2)  A  B  ?  C
3)     B  ?  C  D
4)     B  ?  C
5)  A  B  ?
6)        ?  C  D
7)     B  ?
8)        ?  C
```

Dersom ordene i posisjon A, B, C eller D har flere koder kan regelen ikke brukes. Reglene prøves i rekkefølge 1-8. Hvis en regel løser opp en tvetydighet prøves de andre reglene om igjen.

Eksempel:

```
when            WRB
the             AT
boy's           NN$   NN+BEZ   NN+HVZ
old             JJ
horse           NN    VB
is              BEZ
here            RN
```

I første omgang fins det ingen regel for den første tvetydigheten. Til den neste kan regelen

```
?  BEZ      --> -VB
```

brukes. Nå kan imidlertid regelen

```
?  JJ  NN   --> NN$
```

brukes og begge tvetydigheter er oppløst.

Referanser

Francis, W. Nelson. 1979. Manual of Information to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers. Rev. ed. Providence, RI: Department of Linguistics, Brown University.

Greene, BarbaraB. & Rubin, Gerald M. 1971 Automatic Grammatical Tagging of English, Providence, RI: Department of Linguistics, Brown University

Hofland, Knut & Stig Johansson. 1979. Microfiche concordance of the Lancaster-Oslo/Bergen Corpus. Bergen: NAVFs edb-senter for humanistisk forskning.

Hofland, Knut & Stig Johansson. 1981. Word Frequencies in British and American English. Bergen: NAVFs edb-senter for humanistisk forskning

Johansson, Stig, Leech, Geoffrey N. & Helen Goodluck. 1978. Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers. Engelsk Institutt, Universitetet i Oslo.

Kucera, Henry & W. Nelson F rancis. 1967. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press.

Kucera, Henry & W. Nelson Francis. (under utgivelse) Frequency Analysis of English Usage: Vocabulary and Grammar.

Appendix A

LIST OF TAGS FOR THE BROWN CORPUS

.	., ;, ?, !
(	(
*	NOT, N'T
)	)
-	dash
,	,
:	:
ABL	pre-qualifier (QUITE, RATHER)
ABN	pre-quantifier (HALF, ALL)
ABX	pre-quantifier/double conjunction (BOTH ... AND)
AP	post-determiner (MANY, SEVERAL, NEXT)
AT	article (A, THE, NO)
BE	BE
BED	WERE
BEDZ	WAS
BEG	BEING
BEM	AM
BEN	BEEN
BER	ARE, ART
BEZ	IS
CC	coordinating conjunction (AND, OR)
CD	cardinal numeral (ONE, 1)
CI	conjuction/preposition (AFTER)
CS	subordinating conjunction (IF, ALTHOUGH)
DO	DO
DOD	DID
DOZ	DOES
DT	singular determiner (THIS, THAT)
DTI	singular or plural determiner (SOME, ANY)
DTS	plural determiner (THESE, THOSE)
DTX	determiner/double conjunction (EITHER ... OR)
EX	existential THERE
FW	foreign word (hyphenated to regular tag)
HD	word occurs in headline (hypenated to regular tag)
HV	HAVE
HVD	HAD (past tense)
HVG	HAVING
HVN	HAD (past participle)
IN	preposition (AMONG, BETWEEN)
JJ	adjective
JJR	comparative adjective
JJS	semantically superlative adjective (CHIEF, MAIN)
JJT	morphologically superlative adjective (BIGGEST)
MD	modal auxiliary
NC	cited word (hyphenated to regular tag)
NN	singular or mass noun
NN$	possesive singular noun
NNS	plural noun
NNS$	possesive plural noun
NP	proper noun (may be hyphenated to other tag)

```
NP$  possesive singular proper noun
NPS  plural proper noun
NPS$ possesive plural proper noun
NR   adverbial noun (HOME, TODAY, WEST)
OD   ordinal numeral (FIRST, SECOND)
PN   nominal pronoun (EVERYBODY, NOTHING)
PN$  possesive nominal pronoun
PP$  possesive personal pronoun (MY, OUR)
PP$$ second possesive personal pronoun (MINE, OURS)
PPL  singular reflexive (intensive) pronoun (MYSELF)
PPLS plural reflexive (intensive) pronoun (OURSELVES)
PPO  objective personal pronoun (ME, HIM, IT, THEM)
PPS  3rd. sing. nominative personal pronoun (HE, SHE, IT, ONE)
PPSS other nominative personal pronoun (I, WE, THEY, YOU)
QL   qualifier (VERY, LOTS, FAIRLY)
QLP  post-qualifier (EASY, ENOUGH)
RB   adverb
RBR  comparative adverb
RBT  superlative adverb
RI   adverb/preposition (portmanteau) (ABOVE, ALONG)
RIP  adverb/preposition/particle (portmanteau) (DOWN, IN)
RN   nominal adverb (HERE, THEN, INDOORS)
RP   adverb/particle (BACK, AWAY)
TO   infinitive marker TO
TT   word occurs in title (hyphenated to regular tag)
UH   interjection
VB   verb, base form
VBD  verb, past tense
VBG  verb, present participle, gerund
VBN  verb, past participle
VBZ  verb, 3rd. sing. present
WDT  wh-determiner (WHAT, WHICH)
WP$  possesive wh-pronoun (WHOSE)
WPO  objective wh-pronoun (WHOM, WHICH, THAT)
WPS  nominative wh-pronoun (WHO, WHICH, THAT)
WQL  wh-qualifier (HOW)
WRB  wh-adverb (HOW, WHEN, WHERE)

------------------------

N'T  *
'LL  +MD
'RE  +BER
'VE  +HV
'D   +MD  +HVD  +DOD
'S   $    +BEZ  +HVZ
S'   $
```