Mette-Cathrine Jahr
Stig Johansson
Britisk Institutt
Universitetet i Oslo

GRAMMATISK MERKING AV THE LANCASTER-OSLO/BERGEN CORPUS: ORDKLASSEBESTEMMELSE VED HJELP AV ORDSLUTT

1 Målsetting

I forbindelse med prosjektet "Grammatical Tagging of the Lancaster-Oslo/Bergen Corpus" har vi i Oslo spesielt konsentrert oss om å revidere Greene og Rubins suffiksliste og ordliste for Brown Corpus.[1] Våre reviderte lister tar hensyn til både Brown Corpus og Lancaster-Oslo/Bergen LOB) Corpus, dvs. tilsammen ca. 2 millioner ord løpende tekst. Riktignok er dette et relativt beskjedent materiale i forhold til alt som finnes av engelske tekster, men vi tror likevel at de resultatene vi er kommet frem til, er ganske allmenngyldige.

2 Problemstilling

For den som ønsker å utføre en automatisk grammatisk analyse, byr engelsk på spesielle problemer, både fordi språket inneholder et usedvanlig stort antall homografer, og fordi det så godt som fullstendig mangler distinktive bøyningsendelser. Ikke desto mindre viser Greene og Rubins arbeid (1971) klart at det i stor utstrekning lar seg gjøre å bestemme ordklasser ut fra avledningsendelser og andre typer sluttsekvenser. (Betegnelsene "suffiks" og "endelser" vil her bli brukt i betydningen ordslutt eller sluttsekvens.)

3 Materiale og metode

Da Greene og Rubin (1971) laget sin suffiksliste, brukte de foruten en final-alfabetisk ordliste for Brown Corpus (ca. 50.000 ordtyper), final-alfabetiske ordlister fra Dolby og Resnikoff (1967). Vi har benyttet en lignende fremgangsmåte i arbeidet med å revidere suffikslisten. Følgende materiale ble brukt:

A: En final-alfabetisk ordliste over de tilsammen ca. 75.000 ordtypene (grafiske ord) som finnes i Brown Corpus og LOB Corpus.[2]

B: En frekvensliste over grammatiske koder som forekommer ved hyppige endelser (fra én til fem bokstaver), basert på den grammatisk merkede versjonen av Brown Corpus (heretter kalt suffiks/kode-listen). Noen eksempler:[3]

Grammatisk merking av The Lancaster-Oslo/Bergen Corpus: ordklassebestemmelse ved hjelp av ordslutt
Mette-Cathrine Jahr, Stig Johansson
Proceedings of NODALIDA 1981, pages 125-136

IVE	JJ (= adjektiv)	254
	NN (= substantiv, sg.)	50
	NP (= egennavn)	25
	VB (= verb, infinitiv)	20
	CD (= grunntall)	10
RIVE	VB (= verb, infinitiv)	7
SIVE	JJ (= adjektiv)	65
	NN (= substantiv, sg.)	6
TIVE	JJ (= adjektiv)	182
	NN (= substantiv, sg.)	36
	NP (= egennavn)	16

Nedre frekvensgrense ble satt til 5, dvs. en grammatisk kode måtte opptre sammen med et bestemt suffiks minst fem ganger for å bli tatt inn i listen.

C: Final-alfabetiske ordlister fra Dolby og Resnikoff (1967), som bygger på Shorter Oxford English Dictionary og Merriam Webster New International Dictionary.

D: Forskjellige verker som behandler morfologi og orddannelse i engelsk, spesielt Ljung (1974) og Marchand (1969).

Vi tok vårt utgangspunkt i den final-alfabetiske ordlisten basert på de to korpusene (LOB og Brown) og begynte med å merke ordene etter Greene og Rubins suffiksregler og skille ut alle unntak. Ved å sammenligne antall ord som dekkes av en regel med antall unntak, kunne vi bedømme hvor effektiv hver enkelt regel var. Underveis ble vi oppmerksomme på feil og uoverensstemmelser. Lite effektive regler ble avslørt og nye regelmessigheter ble oppdaget. Som hjelpemidler i arbeidet med å finne frem til nye regelmessige endelser, benyttet vi også listene nevnt under punkt B, C og D ovenfor. Dette skulle gi arbeidet vårt større generell gyldighet.

4 Den reviderte suffikslisten

Resultatet av vårt arbeid er en revidert og utvidet suffiksliste over vel 600 suffikser med tilhørende grammatiske koder. I utgangspunktet mener vi at Greene og Rubins metode er god. Vi har derfor utarbeidet den reviderte suffikslisten etter deres prinsipper. De forandringene vi har gjort, berører ikke selve strukturen i merkeprogrammet deres. Under punkt 6 nedenfor vil vi imidlertid diskutere noen mer radikale endringer som kanskje burde overveies nøyere.

4.1 Stryking av suffiksregler

Ca. 80 av Greene og Rubins suffiksregler er blitt tatt ut av listen. I eksemplene nedenfor er endelsene skrevet forfra på

vanlig måte slik at de er lettere å identifisere og lese:

LB -→ NN: Regelen dekker bare tre ord (Dekalb, Kolb, bulb), hvorav to er egennavn som vi ikke behøver å ta hensyn til i suffikslisten. Det tredje ordet dekkes av en annen eksisterende regel: B -→ NN-VB.

IELD -→ NN-VB: Regelen dekker bare fem ord (shield, windshield, sunshield, wield, yield). Vi lar en annen eksisterende regel behandle dem: LD -→ NN-VB.

ZARD -→ NN: Det er bare fem vanlige ord som slutter på ZARD,[4] hvorav to er unntak fra regelen. De to unntakene settes inn i ordlisten, de øvrige blir ivaretatt av en annen eksisterende regel: RD -→ NN.

RDE -→ NN: Foruten egennavn er det bare ett vanlig ord som dekkes av denne regelen (horde). Vi lager en ny, effektiv suffiksregel: DE -→ NN, som dekker 42 tilfeller og bare gir ett unntak.

UGE -→ NN: Regelen dekker følgende vanlige ord: gauge, refuge, centrifuge, huge, deluge, gouge, rouge. Ett av disse er et adjektiv, de andre kan være både substantiv og verb. Vi setter huge inn i ordlisten og lar de øvrige bli tatt hånd om av en annen eksisterende regel: GE -→ NN-VB.

EPE -→ NN-VB
OUPE -→ NN : Disse reglene behandler bare tre vanlige ord (hvorav ett blir galt kodet): crepe, cantaloupe, troupe. Vi innfører en ny regel: PE -→ NN som dekker 24 ord og gir ett unntak.

GNE -→ NN: Det eneste vanlige ordet som ender på GNE er champagne. Vi innfører en ny regel: NE -→ NN. Denne regelen dekker 46 ord og gir ingen unntak.

EFY -→ VB: To vanlige ord har denne sluttsekvensen, ett av dem et adjektiv: defy og beefy. Vi setter begge i ordlisten.

Noen av Greene og Rubins suffiksregler er høyst merkelige, f.eks. LYE -→ JJ (adjektiv) og LOREN -→ NNS (substantiv, plural). Sekvensen LYE finner vi bare i sitater fra eldre engelsk, og LOREN forekommer bare som et egennavn.

De suffiksene vi har fjernet, dekker svært få ord og ville ofte ha ført til at ord var blitt galt kodet. Som eksemplene ovenfor viser, har vi latt de berørte ordene bli behandlet (1) av andre eksisterende regler, (2) ved å innføre nye og mer effektive regler, eller (3) ved å sette de relevante formene inn i ordlisten.

4.2 Endring av grammatiske koder

I ca. 25 tilfeller der de opprinnelige suffiksreglene ble beholdt,

endret vi de grammatiske kodene. Koder ble fjernet i følgende tilfeller:

		Fjernet kode	Effektivitet [5]
AD	→ NN-JJ	JJ	67/ 8
SIDE	→ JJ-NN	JJ	61/10
LE	→ NN-VB	VB	164/ 2
OLE	→ NN-VB	VB	61/ 5
TIME	→ NN-JJ	JJ	61/ 5
RNE	→ NN-VB	VB	29/ 0
ITE	→ NN-VB-JJ	VB	140/14
H	→ NN-VB	VB	122/12
PH	→ NN-VB	VB	15/ 1
TEN	→ NN-VB	NN	36/ 6
IN	→ NN-VB	VB	400/21
ON	→ NN-VB	VB	393/30
O	→ NN-VB	VB	879/44
ER	→ NN-VB-JJR-RBR	RBR	453/17
IR	→ NN-VB	VB	34/ 4
UENT	→ NN-JJ	NN	15/ 3
EST	→ JJT-RBT	RBT	267/61

Som vi ser av forholdet mellom antall behandlede ord og antall unntak, virker de reviderte reglene ganske godt.[6] Etter mye nøling bestemte vi oss for å stryke RBR (= adverb i komparativ form) og RBT (= adverb i superlativ form) fra kodesettet for ER og EST. Tallene i suffiks/kode-listen taler imidlertid for seg selv:

ER	NN	(= substantiv, sg.)	913
	VB	(= verb, inf.)	123
	JJR	(= adjektiv i komparativ form)	140
	RBR	(= adverb i komparativ form)	24
EST	JJT	(= adjektiv i superlativ form)	143
	RBT	(= adverb i superlativ form)	9

I grammatikker (bl.a. Quirk et al. 1972:294) påpekes det dessuten at det bare er et lite antall adverb som kompareres med ER og EST. Komparativ- og superlativformer på ER og EST som kan være adverb, ble satt inn i ordlisten.

I seks tilfeller føyde vi til koder:

		Tilføyet kode	Effektivitet
AID	→ VBN-VBD	JJ	33/ 4
WARD	→ RB	JJ	44/13
NINE	→ JJ	NN	16/ 1
BORNE	→ VBN	JJ	10/ 0
ESQUE	→ JJ	NN	11/ 1
SIZE	→ NN-JJ	VB	20/ 1

En av våre hovedregler har vært ikke å sette til nye koder fordi dette innebærer at reglene blir mindre presise. Vi har i stedet foretrukket å spesifisere unntak i ordlisten. For suffiksreglene nevnt ovenfor, ville antall unntak imidlertid ha blitt uakseptabelt høyt hvis vi ikke hadde satt inn de nye kodene. I to tilfel-

ler var vi i sterk tvil om hvorvidt vi skulle føye til nye koder eller ikke:

ED	-→ VBN-VBD	+JJ?
ING	-→ VBG	+NN-JJ?

Siden ord på ED ofte er adjektiver og de på ING ofte adjektiver eller substantiver, var det fristende å sette til de kodene vi har ført opp ovenfor med spørsmålstegn. Når vi allikevel bestemte oss for ikke å gjøre det her, var det dels på grunn av at verbformene av slike ord forekommer langt hyppigere, dels fordi vi ikke ønsket å ødelegge mulighetene for en sammenligning mellom LOB og Brown på dette punktet.

I følgende tre tilfeller endret vi de grammatiske kodene helt:

		Ny kode	Effektivitet
UND	-→ JJ	NN	16/3
EDE	-→ NN	VB	13/2
WHERE	-→ NN	RB	7/1

Vi følte oss også fristet til å endre koden til NP i en del tilfeller:

		Effektivitet med NP-kode
WE	-→ NN-VB	22/ 3
I	-→ NNS	559/58
Z	-→ NN-VB	139/11
HR	-→ NN	8/ 0

Langt de fleste ord med disse sluttsekvensene er egennavn. Men siden egennavn behandles gjennom en spesialrutine i merkeprogrammet (se Hoflands artikkel), bestemte vi oss for ikke å endre kodene. To av de opprinnelige suffiksreglene ble sløyfet (for WE og HR), og ord som ikke er egennavn, ble satt inn i ordlisten. Regelen I -→ NNS ble beholdt, siden pluralformer av substantiver utgjør en ganske stor gruppe av ord på I. Vi beholdt også regelen Z -→ NN-VB. Denne regelen gir ingen unntak selv om den behandler svært få vanlige ord.

Meningen er at rekkefølgen av kodene skal gi uttrykk for den relative frekvensen, dvs. at den hyppigste og mest sannsynlige koden kommer først. Vi endret derfor rekkefølgen på kodene i følgende tilfeller:

		Rettet til
AC	-→ JJ-NN	NN-JJ
ID	-→ NN-JJ	JJ-NN
END	-→ NN-VB	VB-NN
ISE	-→ NN-VB	VB-NN
WISE	-→ JJ-RB	RB-JJ
ERSE	-→ NN-VB-JJ	VB-JJ-NN
ETE	-→ NN-VB-JJ	JJ-NN-VB
TIVE	-→ NN-JJ	JJ-NN

		Rettet til
SH	-→ VB-NN	NN-VB
ASH	-→ VB-NN	NN-VB
IAN	-→ NN-JJ	JJ-NN
HUMAN	-→ NN-JJ	JJ-NN
LIER	-→ NN-JJR	JJR-NN
NCT	-→ NN-JJ	JJ-NN
ANT	-→ NN-JJ	JJ-NN
NENT	-→ NN-JJ	JJ-NN
RENT	-→ NN-JJ	JJ-NN
TENT	-→ NN-JJ	JJ-NN
LY	-→ JJ-RB	RB-JJ
ERY	-→ JJ-NN	NN-JJ
ARRY	-→ NN-VB	VB-NN

For å bestemme rekkefølgen på kodene, brukte vi suffiks/kode-listen for Brown Corpus og den merkede final-alfabetiske ord-listen for både LOB Corpus og Brown Corpus (jfr. punkt 3 A og B). Den førstnevnte listen er ikke tilstrekkelig i seg selv fordi den bare tar hensyn til Brown Corpus, og fordi den gir frekvensene for hver sluttsekvens uten hensyn til suffiksreglene, slik at alle ord på f.eks. LERY og NERY er inkludert i "statistikken" for ord som slutter på ERY. I suffikslisten viser "statistikken" bare de forekomstene på ERY som ikke allerede er behandlet av reglene for LERY og NERY.

4.3 Innføring av nye suffiksregler

Den største endringen i forhold til Greene og Rubin er at vi har innført ca. 240 nye suffiksregler. I eksemplene nedenfor er nye suffikser understreket:

		Effektivitet
IC	-→ JJ-NN	256/ 1
RIC	-→ JJ	81/ 7
ISTIC	-→ JJ	81/ 1
D	-→ NN-VB	165/23
HOOD	-→ NN	23/ 1
NE	-→ NN	46/ 0
INE	-→ NN-VB	222/20
NINE	-→ NN-JJ	16/ 1
RINE	-→ NN-JJ	27/ 1
TINE	-→ NN-JJ	38/ 1
ZINE	-→ NN	8/ 0
TE	-→ NN	70/ 1
ATE	-→ VB-NN-JJ	266/ 2
CATE	-→ VB-NN	34/ 4
DATE	-→ VB-NN	14/ 1
GATE	-→ VB-NN	46/ 2
PHATE	-→ NN	9/ 2
IATE	-→ VB-NN	34/ 6
LATE	-→ VB-NN	68/ 4
TATE	-→ VB-NN	40/ 0
VATE	-→ VB	12/ 1

		Effektivitet
AL	-> JJ-NN	811/24
ICAL	-> JJ	243/ 6
IONAL	-> JJ	106/ 6
UR	-> NN-VB	38/ 0
EUR	-> NN	19/ 0

De fleste av de nye suffiksreglene gir en mer presis ordklasse-angivelse. For RIC og ISTIC, for eksempel, trenger vi bare én kode, mens de ville få to koder hvis de skulle behandles av IC, som i Greene og Rubin. Innføringen av NE og TE illustrerer en annen endring i forhold til Greene og Rubin. Alle ord som slutter på NE og TE må de enten føre opp i ordlisten eller merke NN-VB-JJ, dvs. det kodesettet som blir gitt til alle ord som ikke dekkes av listene eller andre rutiner i merkeprogrammet.

En del nye suffikser er blitt tatt inn i listen som følge av at vi har forsøkt å være mer konsekvente enn Greene og Rubin (1971: 30) når det gjelder å gi så mange ord som mulig en entydig grammatisk kode så tidlig som mulig, selv om dette ikke nødvendigvis betyr en mer nøyaktig ordklasse-angivelse. Vi kan gi noen eksempler som illustrerer dette (nye sekvenser er understreket):

			Effektivitet
ELY	-> RB		134/16
ATELY	-> RB		42/ 2
IVELY	-> RB		90/ 1
ALLY	-> RB		98/ 3
CALLY	-> RB		191/ 0
NALLY	-> RB		54/ 0
RALLY	-> RB		22/ 1
UALLY	-> RB		27/ 0
IST	-> NN		58/ 7
CIST	-> NN		11/ 0
OGIST	-> NN	(G&R: GIST)	23/ 0
LIST	-> NN		55/ 2
MIST	-> NN		13/ 1
NIST	-> NN		53/ 0
RIST	-> NN		33/ 2
OGY	-> NN		6/ 0
OLOGY	-> NN		54/ 0

Nye suffiksregler ble inkludert hvis de er svært effektive og/eller hvis de representerer produktive endelser. Vi kan ikke påstå at vi har vært hundre prosent konsekvente. Noen av våre suffiksregler kunne godt utelates og andre settes inn, men vi har ikke sett på dette som så forferdelig viktig, siden slike endringer ikke har noen innvirkning på valget av grammatisk kode.

5 Den reviderte ordlisten

Vår reviderte suffiksliste er supplert med en ny ordliste på noe under 5000 ord med tilhørende grammatiske koder. Ordlisten inneholder unntak fra suffiksreglene såvel som alle ord i LOB Corpus med en frekvens på 50 eller høyere. Vi har sett bort fra følgende unntak: utenlandske ord, arkaiske former, forkortelser og ord med unormale stavemåter. Videre har vi ikke tatt hensyn til egennavn og ord med bindestrek, fordi disse behandles ved hjelp av spesielle rutiner i merkeprogrammet.

I arbeidet med å skille ut unntak og bestemme grammatiske koder har vi hovedsakelig holdt oss til Longman Dictionary of Contemporary English (1978-utgaven), selv om vi også har brukt andre oppslagsverker, spesielt A Grammar of Contemporary English (1972).

For å øke nytten av den reviderte ordlisten, har vi i tillegg tatt med alle former fra Greene og Rubins ordliste som enten er unntak fra våre suffikslister, eller har en frekvens på 50 eller høyere (i Brown Corpus). Dette skulle gjøre arbeidet vårt mer anvendelig, idet det går ut over grensene for vårt primære siktemål, nemlig en grammatisk merking av LOB Corpus.

6 Diskusjon

Våre reviderte ord- og suffikslister tar hånd om en betydelig del av de tilsammen ca. 75.000 ordtypene i LOB Corpus og Brown Corpus. Hvis antall ord som behandles av suffiksreglene legges sammen med antall ord i ordlisten, kommer vi opp i over 50.000. De aller fleste av disse får bare én grammatisk kode. Imidlertid er det vanskelig å si helt nøyaktig hvor effektive de nye listene er. Tallet 50.000 er for så vidt for høyt, siden det inkluderer egennavn og ord med bindestrek, som behandles av spesialrutiner i merkeprogrammet (rutinen for ord med bindestrek gjør riktignok også bruk av listene).[7] På den annen side blir et stort antall av de resterende ca. 25.000 ordtypene tatt hånd om av andre spesialrutiner i programmet. Dette gjelder særlig grunntall og ordenstall, ord med apostrof og de fleste ord på S (hvor spesialrutinen også anvender ordlisten).

Ikke desto mindre, hva de nøyaktige tallene for "effektivitet" enn måtte være, er det helt klart at de er svært høye. Resultatene blir enda mer imponerende hvis vi, i stedet for å se på antall ordtyper som behandles, ser på det totale antall løpende ord i de to korpusene som blir tilfredsstillende merket ved hjelp av listene. Dette har vi ikke regnet ut, men tallet må være enormt høyt, fordi ordlisten omfatter alle høyfrekvente ord i begge korpusene, og vi vet at disse utgjør en meget stor del av en løpende tekst.

Det er imidlertid ikke nok at de reviderte listene kan behandle alle ord i de to korpusene. For at de skal kunne ha større allmenn interesse, må de også kunne anvendes på andre engelske tekster. Greene og Rubin (1971:41) hevder at det er "almost certain that any 'new' word contained in a sample of present-day

American English will be given its correct tag(s) by matching with the Suffix List". Vi bestemte oss for å teste vår reviderte suffiksliste på et utvalg av nye ord. Femten sider av A Dictionary of New English (1963-72) ble vilkårlig utvalgt, og alle nye former på disse sidene ble registrert. Egennavn, ord med bindestrek og forkortelser ble holdt utenfor testen. Videre unnlot vi å ta med tilfeller der gamle, etablerte ord hadde fått nye betydninger, med mindre dette falt sammen med nye grammatiske funksjoner. Av de 94 ordene vi registrerte, fikk 52 én enkelt, korrekt kode, 33 fikk to koder hvorav én var den riktige. Ett ord fikk tre koder inklusive den riktige, et annet ord ble ikke behandlet i det hele tatt, og syv ord ble galt kodet.[8]

Dette fører oss over til en vurdering av visse svakheter ved suffikslisten vår, og mulige forbedringer av merkeprogrammet.

Fire av de syv ordene som ble galt kodet, var adjektiver med prefikset multi-. Selv om prefikser vanligvis ikke er særlig pålitelige som ordklasse-indikatorer, er det åpenbart noen som ganske regelmessig opptrer ved spesielle ordklasser. Følgende prefikser innleder for eksempel bare substantiver og adjektiver:

micro-	pseudo-
multi-	semi-
non-	ultra-
proto-	vice-

En av Greene og Rubins spesialrutiner ser på en kombinasjon av et prefiks og et suffiks, nemlig UN...ED. Det måtte være mulig å inkorporere en prefiksrutine ved sekvenser som er typiske for spesielle ordklasser.

To av de nye ordene som ble galt kodet, er eksempler på substantiver brukt som verb (collage og network). Det er ikke lett å sikre seg riktig koding av slike tilfeller bare ved hjelp av endringer i suffikslisten eller andre deler av merkeprogrammet. Vi må innrømme at det alltid vil være en liten gruppe ord som ikke kan behandles riktig. Dette er prisen vi må betale for å oppnå en mer eksakt spesifisering i de aller fleste tilfellene.

En alvorlig innvending mot Greene og Rubins suffiksliste, såvel som vår egen reviderte liste, er at den ikke gjør bruk av visse opplagte regelmessigheter i engelsk orddannelse. Vi vet for eksempel at ord som ender på ER er komparativformer av adjektiver hvis resten av ordet er et adjektiv (old-er), og at de er substantiver hvis den øvrige delen av ordet er et verb (speak-er), etc. Greene og Rubin hadde en regel som gav disse ordene fire koder: ER -→ NN-VB-JJR-RBR. Dette betyr at store grupper av vanlige ord får flere koder, noe som vil føre til problemer på et senere stadium når kontekstreglene skal tre i funksjon. Det er også andre vanlige suffikser som ikke blir tilfredsstillende behandlet:

ED -→ VBN-VBD: Man utnytter ikke det faktum at ED-former er adjektiver når det ikke er et verb som går forut for ED (wick-ed).

EN --> JJ-NN: Vi klarer ikke å fange opp at EN er en regelmessig adjektiv-endelse hvis det står et substantiv foran (wood-en, earth-en) og like regelmessig en verb-endelse hvis det kommer et adjektiv foran (black-en, stiff-en).

ISH --> JJ-VB: Man kan faktisk si at hvis det som står foran ISH er et ord, er kombinasjonen et adjektiv (seven-ish, grey-ish, boy-ish).

LY --> RB-JJ: Vi går glipp av muligheten til å trekke den opplagte slutning at et adjektiv fulgt av LY er et adverb.

Y --> JJ-NN: Vi kan ikke få frem at et substantiv fulgt av Y normalt er et adjektiv (beef-y, hill-y).

Ved alle disse vanlige, produktive endelsene fører våre regler til unødig tvetydighet, selv om tvetydigheten i mange tilfeller reduseres ved at vi innfører lengre sekvenser med entydige koder (spesielt når det gjelder ord på LY). Det ville sannsynligvis være mye bedre å innføre flere avkortingsrutiner som svarer til den for ord som slutter på S (se Hoflands artikkel).

Vi har kommet med noen forslag til hvordan Greene og Rubins merkeprogram skulle kunne forbedres ved å innføre spesialrutiner som innebærer at programmet ser på prefikser og kutter av suffikser. Selv om denne delen av merkeprogrammet muligens blir noe mer komplisert enn nå,[9] vil dette delvis kunne kompenseres ved at man kan eliminere noen av de eksisterende spesialrutinene (jfr. Hoflands artikkel). Grunntall og ordenstall kunne for eksempel behandles av et lite antall suffiksregler, og mange vanlige former med apostrof kunne bare settes inn i ordlisten. Legg også merke til at spesialrutinen for ord som UN...ED ville bli overflødig hvis ED ble behandlet på en mer generell måte. Videre ville lengden på suffikslisten kunne reduseres betraktelig hvis suffiksavkorting ble brukt i større utstrekning.

Til slutt vil vi nevne at det er en mengde suffikser som nesten utelukkende forekommer i egennavn (BURGH, SHIRE, FORD, SKI, etc.). Det er klart at et stort antall egennavn kunne behandles ved at man innfører NP-suffiksregler. Dette ville redusere behovet for pre-editing (som den nåværende rutinen forutsetter).

7 Konklusjon

Selv om det finnes problemer i forbindelse med suffikslisten (og mulige forbedringer er blitt foreslått), må vår konklusjon bli at den i det store og hele fungerer meget bra.[10] Vi har sett at listen gir tilfredsstillende behandling av ordene i de to korpusene såvel som av de "nye" ordene i utdragene fra A Dictionary of New English. Kan dette bety at det er større overensstemmelse mellom ordform og ordklasse i engelsk enn det vi hittil har antatt? Det er helt opplagt at vi finner en ganske stor grad av regelmessighet så snart vi beveger oss utenfor den sentrale kjernen av høyfrekvente, først og fremst germanske, ord. Det er interessant å observere at denne regelmessigheten ikke begrenser

seg til suffikser i lingvistisk forstand, men også omfatter ordslutt i sin alminnelighet. Blant de "nye" ordene som fikk én eneste og riktig kode, finner vi former som duende, duka og dumdum. Regelmessigheter av denne art kunne bare oppdages ved hjelp av datamaskin og bør uten tvil utnyttes maksimalt i automatisk grammatisk analyse.

FOTNOTER

1 Knut Hoflands artikkel (i denne publikasjonen) om grammatisk merking av LOB Corpus tar også for seg Greene og Rubins "Automatic Grammatical Tagging of English" (1971) basert på Brown Corpus. Vi forutsetter derfor at dette er kjent stoff og henviser heretter til Hoflands artikkel når det gjelder spesielle punkter i Greene og Rubins merkeprogram.

2 Listene nevnt under punkt A og B er laget ved NAVF's EDB-senter for humanistisk forskning i Bergen, Norge. Programmerer: Knut Hofland.

3 Suffikser og grammatiske koder vil bli skrevet med store bokstaver, slik at det er lettere å skille dem ut fra den øvrige teksten.

4 Med "vanlige ord" mener vi enkeltord (uten bindestrek) med normal stavemåte, unntatt egennavn og forkortelser. Egennavn og ord med bindestrek behandles ved egne rutiner i programmet (se Hoflands artikkel).

5 "Statistikken" som er oppført ved hver regel i den reviderte suffikslisten angir hvor "effektiv" hver regel er. Tallene i venstre kolonne angir det totale antall ord med en bestemt sluttsekvens (som ikke allerede er behandlet av regler for lengre sekvenser), mens tallet i høyre kolonne angir antall unntak fra regelen.

6 De opprinnelige suffiksreglene ville også ha gitt en god del unntak.

7 Inkludert i dette tallet er også ca. 10.000 forekomster av ord på ED og ING som vi mener ikke blir tilfredsstillende kodet.

8 Resultatene av denne testen fikk oss til å foreta noen mindre endringer i suffikslisten, slik at antall galt kodede ord dermed er blitt redusert fra syv til fem (4 adjektiver med prefikset multi- samt network brukt som verb). En ny suffiksregel (É -→ NN-JJ) tar seg nå av det ordet som ikke ble behandlet i det hele tatt.

9 En del av disse rutinene vil muligens forutsette en større ordliste.

10 Den fullstendige suffikslisten vil bli publisert i Johansson og Jahr (under utgivelse).

REFERANSER

Barnhart, Clarence L., Steinmetz, Sol og Robert K. Barnhart. 1973. A Dictionary of New English 1963-1972. London: Longman.

Dolby, J.L. og H.L. Resnikoff. 1967. The English Word Speculum. Bind III og V. Den Haag: Mouton.

Greene, Barbara B. og Gerald M. Rubin. 1971. Automatic Grammatical Tagging of English. Providence, R.I.: Department of English, Brown University.

Johansson, Stig og Mette-Cathrine Jahr. "Grammatical Tagging of the Lancaster-Oslo/Bergen Corpus: Predicting Word Class from Word Endings". I Stig Johansson, utg., Computer Corpora in English Language Research. NAVF's EDB-senter for humanistisk forskning, Bergen. (under utgivelse)

Ljung, Magnus. 1974. A Frequency Dictionary of English Morphemes. Data linguistica 9, Universitetet i Göteborg. Stockholm: AWE/Gebers.

Longman Dictionary of Contemporary English. 1978. London: Longman.

Marchand, Hans. 1969. The Categories and Types of Present-Day English Word-Formation. Annen utgave. München: C.H. Beck'sche Verlagsbuchhandlung.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey N. og Jan Svartvik. 1972. A Grammar of Contemporary English. London: Longman.