

AUTOMATISK MORFOLOGISK ANALYSE AF DANSKE TEKSTER

Af Thomas Bilgram og Hans Arndt (Aarhus Universitet)

Vi skal i det følgende præsentere en automatisk morfologisk analyse af danske tekster. Den forudsætter en del teknik, som vi kun kort vil komme ind på. Hovedsagen er beskrivelsen af de principper som den elektroniske behandling hviler på, og den form som den morfologiske analyse får.

Den morfologiske analyse indgår som en del af et projekt hvis formål er at udvikle en automatisk syntaktisk analyse ('parsing') af naturlige danske skriftsprogstekster¹⁾. Den morfologiske analyse er en forudsætning for den syntaktiske, i den (ret trivielle) forstand at morfologien (og de grammatiske kategoriseringer der kan udledes heraf) tillader en delvis bestemmelse af ordenes syntaktiske funktion i teksten. Det betyder at en morfologisk analyse kan give et kvalificeret input til en syntaktisk analyse²⁾. Og eftersom den morfologiske analyse er simplere end den syntaktiske, er den det naturlige udgangspunkt.

Den automatiske morfologiske analyse forudsætter 1) en teoretisk og programmeringsteknisk basis, 2) et struktureret og markeret leksikon, 3) et sæt regler for fleksiv, derivativ og kompositionel orddannelse (morfotaks), og 4) et sæt regler for allomorfisk (morfofonemisk) variation. Førstnævnte vil vi anse som universel, mens 2-4 må være specielle for dansk. Vi vil nedenfor redegøre for disse fire komponenter.

1 Basis

Inspirationen for vores arbejde er dels teoretisk, dels praktisk. Den teoretiske inspiration udspringer af en erkendelse af at sproglig og tekstuel analyse styrkes

1) Det er tanken at analysen skal udvides til også at omfatte mundtligt sprog; men det har vi ikke for nuværende empirisk grundlag for.

2) Det er klart at sprogenes typologi gør en væsentlig forskel. Således giver den morfologiske analyse af sprog som tysk, latin eller finsk et stærkere grundlag for den syntaktiske parsing end mere isolerende sprog som dansk eller engelsk. Men selv ved analysen af sidstnævnte rummer morfologien nøgler, som det vil være spild ikke at udnytte i den syntaktiske analyse.

væsentligt af adgang til empirisk eksemplifikation i (korpora af) naturlige tekster, og af at 'traditionelle' alfanumeriske søgeprocedurer kun nødtørftigt tilgodeser empirisk søgning på syntaktisk (og evt. pragmatisk) analyseniveau. For at opnå sidstnævnte er der behov for en forbehandling af korpora: en automatisk markering (tagging) af ordenes syntaktiske funktion.

Den praktiske inspiration finder vi i det faktum at der allerede foregår arbejde med automatisk tagging (ordbeskrivelse baseret på morfologisk analyse og syntaktisk parsing) med andre sprog, jf. Garside et al. 1987 (Lancaster-Leeds-projektet) og Karlsson 1992 (Helsinki-modellen). Vi har valgt sidstnævnte som vores udgangspunkt.

Til den morfologiske analyse kalkuleres der i Helsinki-modellen med et rodleksikon bestående af 40-50.000 opslagsord, dækkende noget færre leksemer, samt diverse 'kontinuations'-leksika, dvs. små subleksika, indeholdende fleksive og derivative affikser. Reglerne for kontinuation (fra én morfem-type til en anden) beskriver sprogets morfotaks; modellen rummer desuden mulighed for sammenknytning af rodleksemer (rekursiv kontinuation), til analyse af komposita. Endelig rummer modellen en programmerings-mekanisme til genkendelse af allomorfiske varianter. Specielt på grund af sidstnævnte kendes Helsinki-modellen som en 'to-niveau-model', idet allomorfisk variation betragtes som en overflade-repræsentation af en underliggende leksikal repræsentation³⁾.

Et eksempel vil vise hvorledes modellen behandler den morfotaktiske kontinuation. Ordet *kulturen* vil få følgende fire analyser: 1) 'kult•u|ren', 2) 'kult•uren', 3) 'kul•tur|en' og 4) 'kultur|en' ('•' står for kompositumgrænse, '|' for morfemgrænse).

Analyse 1 ('kult•u|ren') tillades af at de tre morfemer findes i leksikon, af at sammensætningen 'substantiv•adjektiv' er en tilladt kontinuation, og af at *u|* er et præfiks der tillader et adjektiv som kontinuation. Analyse 2 ('kult•uren') tillades af at ordets to komponenter findes i leksikon, og af at sammensætningen

3) En omhyggelig og kritisk gennemgang af to-niveau-modellen kan findes hos Hansen 1989.

substantiv•adjektiv' er en tilladt kontinuation.⁴⁾ Analyse 3 ('kul•tur|en') tillades af at de tre morfemer findes i leksikon, at sammensætningen 'substantiv•substantiv' er en tilladt kontinuation (rekursivitet⁵⁾) og af at bestemmheds-endelsen |en er en tilladt kontinuation efter substantivet *tur*. Analyse 4 ('kultur|en') tillades af at de to morfemer findes i leksikon, og af at bestemmheds-endelsen |en er en tilladt kontinuation efter substantivet *kultur*.

Et andet eksempel vil vise hvorledes modellens allomorfiske komponent klarer en lemmatisering af ordet *fædre*. Analysen giver 'fa•dAr|e'; én to-niveau-regel tillader at den leksikale læsning af 'a' på overfladen læses som -æ-, når leksemet følges af en pluralis-endelse. En anden tillader at 'A' læses som zero foran pluralis-endelse, og som e foran singularis-endelse, ordafslutning eller kompositumskontinuation (jf. *faderen* og *faderbinding*, ikke **fadren* og **fadrbinding*).

De programmeringstekniske principper bag analysen er udviklet af Kimmo Koskeniemi, Helsinki Universitet. Den er omsat til PC-, Macintosh- og UNIX-programmer af Evan L. Antworth (1990). Vi har brugt Macintosh-versionen, da en almindelig pc'er har en begrænsning i størrelsen af leksikon.

For tiden kræver programmet med den aktuelle størrelse leksikon over 6 Mb. RAM, og kan således bedst køres på en maskine med 8 Mb. RAM eller derover.

Vi skal nedenfor gøre nærmere rede for hvilke regler det har været nødvendigt at opstille for at analysere danske ord i henseende til fleksion, derivation, komposition og allomorfi. Som nævnt ses denne analyse som en forudsætning for en syntaktisk analyse; sammenhængen hertil er de tags vi ved morfologiens hjælp kan hæfte på de enkelte tekstord. De vil blive nærmere omtalt til sidst. Om principperne for taggingen, se Arndt (in press).

4) Når analyse 1 og 2 begge forekommer, er det fordi almindeligt forekommende ord præfigerede med 'u' findes i leksikon som lemmata, og desuden accepteres som orddannelser i henhold til kontinuationsreglerne. Med netop præfikset 'u' er det sandsynligt at det i den sidste ende viser sig at være så overgenererende, at dette præfiks må slettes af leksikon.

5) Også kombinationen 'substantiv•adjektiv' er rekursiv, for så vidt som der er tale om rekursion til rodleksikon — men altså til en anden del af rodleksikon.

2 Leksikon

Det er en grundlæggende ide i Helsinki-modellen at den automatiske tagging forudsætter et fuldt dækkende leksikon⁶⁾. Der er ikke mulighed for en ren morfologisk bestemmelse af et ord, hvis ikke dets komponenter findes i leksikon.

Det er imidlertid af flere grunde upraktisk at arbejde med ét leksikon der er stort nok til alle behov, og strategien er derfor at der opbygges et basisleksikon, og at der sidenhen opbygges en række specialleksika, til brug ved tagging af tekster der er specielt fag- eller typebestemte. Det er klart at en sådan strategi dels kræver løbende justering af leksikon⁷⁾, dels også muliggør en tekst- eller genre-bestemmelse, når flere typer leksika er udbygget. Vi koncentrerer os foreløbig om basisleksikon, som skal kunne klare tekster som f.eks. dem der indgår i Korpus DK87-90 (Bergenholtz 1991).

Basisleksikon skal yderligere indeholde almindeligt forekommende derivede og komponerede ord - behovet herfor vil fremgå nedenfor i sektion 6.1. Endelig skal det indeholde almindeligt forekommende egen- og sted-navne.

Bindestreks-ellipse (som vi netop har brugt) samt andre former for tegnsætning (f.eks. ved forkortelser) volder forskellige problemer, som bl.a. kan løses ved en 'pre-processing' der regulariserer variante tegnbrug og stavemåder⁸⁾.

Der er foreløbig udfærdiget et basisleksikon på ca. 44.000 opslagsord (jf. Bilgram 1992), dækkende noget færre leksemer, da en del af opslagsordene er komposita. Dette basisleksikon (herefter leksikon) er oprettet på basis af Dansk Sprognævns Retskrivningsordbog (RO). Det indeholder, i en tillempet form, de oplysninger om ordklasse og fleksion som findes i RO. For ord uden morfologi markeres mulige (syntaktisk bestemte) ordklasser, hvilket muliggør en kvalificeret forudsigelse af ordets syntaktiske funktion. RO indeholder desuden en angivelse

6) Til gengæld er oprettelsen af leksikon en éngangsforeteelse — det kan bruges til tagging af et ubegrænset antal tekster.

7) Det er nødvendigt i enhver form for computerbehandling af leksikale og leksikografiske forhold, jf. Sinclair 1991.

8) Egentlige 'fejl'stævninger (i henhold til RO) kan systemet ikke klare — de giver fejlmelding hvis resultatet ikke findes, og fejlanalyse hvis der er tale om et muligt ord. På meget langt sigt skal systemet programmeres til at opfange sidstnævnte ved den syntaktiske analyse hvis resultatet er syntaktisk uacceptabelt, og til generelt at ledsage fejlmeldinger med et rettellesforslag.

af 'fugevokal' eller 'fuge-s' ved kompositum-dannelse; denne oplysning er ikke medtaget i leksikon, men ligger som en del af de morfotaktiske regler.

Vi har haft mulighed for at bruge RO i en maskinlæsbar form og har ved hjælp af et HyperCard-program udtaget alle de regelmæssige, og enskodede, informationer. Restgruppen på ca. 10.000 uregelmæssige ord er delvis manuelt indført i leksikon.

3 Morfologi

Hvert opslagsord i leksikon er kodet for sin morfotaktisk mulige continuation. Det betyder at det er muligt for den automatiske tagging at søge efter leksikale korrespondenser for et tekstord efter følgende skema:

præfiks* → rod* → derivationssuffiks* → fleksionssuffiks

Hvert trin står for en søgning i et (sub)leksikon. De med * markerede positioner angiver rekursion, dvs. muligheden for at gå tilbage til udgangspositionen. Alle trin i søgningen er optionelle, dog skal et ord have mindst én rod; men når den er registreret, kan man godt f.eks. have derivationssuffiks+derivationssuffiks (f.eks. *smør|bar|hød*). Visse fleksionssuffikser kan følge efter andre (f.eks. *bil|er|ne|s*); bortset herfra repræsenterer fleksionssuffigering afslutningen på en ordanalyse.

To eksempler: ordet *ren* vil kun kunne findes i rod-leksikonet; ordet *antiforureningskampagnens* vil (blandt andet) afstedkomme søgning (med resultat) på 'præfiks-præfiks-præfiks-rod-derivationssuffiks-rod-fleksionssuffiks-fleksionssuffiks': *anti|for|u|ren|ings|kampagne|n|s*, men det vil også få analysen 'præfiks-rod-derivationssuffiks-rod-fleksionssuffiks-fleksionssuffiks': *anti|foru|ren|ings|kampagne|n|s*, fordi verbet *forurene* også findes som opslagsord.

3.1 Fleksion

I kontinuationskoden for et substantiv indgår en markering af genus og numerus:

<u>genus</u>	<u>pluralisform</u>
FÆL (fælleskøn)	PLR (-er)
INT (intetkøn)	PLE (-e)
FI (fælles- & intetkøn)	PLS (-s)
	PLØ (ø)

PLEE (-e)

Kodningen for genus anviser continuationen *-en* (FÆL) eller *-et* (INT) eller en kombination (FI, f.eks. for *katalogen/kataloget*).

Kodningen for pluralis anviser de rette pluralis-suffikser. Også de kan kombineres, da nogle ord varierer mellem flere pluralismorfemer. F.eks. forekommer koden FÆLXPLRE for ordet *minke*, der kan hedde både *minke* eller *minke* i pluralis; og INTXPLEØ for ordet *tog*, der kan hedde både *toge* og *tog* i pluralis. ('X' bruges som adskillelse mellem de to kodeelementer.) Alle suffikser, såvel som roden, har genitivskontinuationen *-s*.

Pluralis-suffikser har mulighed for continuationen *-ne*, med en særrregel for PLEE, der dækker de ord hvor pluralis-bestemtheds morfemet *-ne* sættes direkte på stammen (f.eks. *jægere—jægerne, danskere—danskerne*). Der er oftest tale om fælleskøns-substantiver af nomen-agentis-typen eller nationalitetsbetegnelser, men der findes desuden (i henhold til RO) 14 ord i intetkøn, f.eks. *kadaver* og *pulver*, og et par ord hvor kønnet kan være begge dele, nemlig *naver* og *polyester*. Kodningen PLEE findes ikke kombineret med andre.

Den morfologiske analyse af substantiverne muliggør en tagging af det enkelte tekstord der oplyser genus, numerus og kasus, som alle kan være signifikante for den syntaktiske analyse.

Verbernes continuation er kodet for synkope, samt for præteritums- eller præteritum-participiums-form:

<u>synkope</u>	<u>præteritum eller præteritum participium</u>
NOR (uden synkope)	EDE (præt. <i>-ede</i>)
SYN (med synkope)	TE (præt. <i>-te</i>)
	MP (samme stamme i præsens og i præteritum participium)
	UP (forskellig stamme i præsens og i præteritum participium)

Fælles for verber kodet med SYN er, at der sker bortfald af *-e-* i infinitiv, samt præsens aktiv og passiv (jf. *slå—slår—slås* og *se—ser—ses*). Gruppen indeholder ialt 47 simplex-ord, næsten alle enstavelsesord der ender på vokal, som dog giver over hundrede leksikale opslagsord når komposita medtages.

De første to af de fire præteritums- eller participiums-angivelser karakteriserer de 'regelmæssige' verber efter deres præteritumsformer, *-ede* og *-te* (jf. *hentedede* og *kørte*). De sidste to deler de 'stærke' verber i to typer, jf. *drage*, *draget* (MP), men *drikke*, *drukket* (UP). Andre uregelmæssige verber (som *flyve*, *fløj*, *fløjet*) behandles individuelt.

Den morfologiske analyse af verberne muliggør en tagging af det enkelte tekstord for finithed, modus og participiumsform, der alle kan være signifikante for den syntaktiske analyse.

Adjektivernes kontinuation indholder oplysning om *-e*, *-t* og *-ø*-morfemernes mulige forekomst. Der findes 6 typer:

ADJ1	<i>-e</i>	<i>-t</i>	<i>-ø</i>	Eks.: <i>dejlig</i>
ADJ2	<i>-e</i>	(<i>-t</i>)	<i>-ø</i>	Eks.: <i>rask</i> , <i>frisk</i>
ADJ3	<i>-e</i>		<i>-ø</i>	Eks.: <i>naivistisk</i> , <i>dansk</i> , <i>sort</i>
ADJ4	(<i>-e</i>)	<i>-t</i>	<i>-ø</i>	Eks.: <i>ny</i> , <i>fri</i>
ADJ5		<i>-t</i>	<i>-ø</i>	Eks.: <i>blå</i> , <i>rå</i>
ADJ6			<i>-ø</i>	Eks.: <i>ringe</i>

Som det fremgår, kan morfemerne *-e* og *-t* af fonotaktiske grunde forsvinde, eller være optionelle. Men adjektiver med kontinuationen ADJ1 er langt den største gruppe, og ADJ4-6 gælder kun ganske få ord.

Det er signifikant for den syntaktiske analyse at forekomsten af et bestemt suffiks begrænser det pågældende adjektivs kongruens-mulighed, således (for ADJ1) *-t* : ubestemt, singularis, intetkøn (eller adverbium); *-ø* : ubestemt, singularis, fælleskøn; *-e* : bestemt/ubestemt, singularis/pluralis, intetkøn/fælleskøn. Disse oplysninger indgår i taggingen.

Samtlige adjektiver kan kompareres med morfemerne *-ere* og *-est*. En lille gruppe, alle endende på *-ig*, danner superlativ med *-st*. Da de alle er af typen ADJ1, har de fået kontinuationskoden ADJ1XST. Uregelmæssigt komparerede adjektiver (som *god*, *bedre*, *bedst*) behandles individuelt.

Pronominer betragtes i almindelighed som flekterbare ord. Deres fleksion er imidlertid normalt ikke morfotaktisk analysérbar, men markeres individuelt i leksikon. Deres tagging mht. genus, numerus, kasus og syntaktisk type er signifikant for den syntaktiske analyse.

3.2 Derivation

Afgrænsningen af derivation over for fleksion på den ene side og komposition på den anden kan være kontroversiel. For den morfotaktiske kontinuationsanalyse er spørgsmålet kun vigtigt for så vidt som det skal afgøres hvad et affiks (eller en rod) kan følges af.

Vi må følgelig anse det adverbielle *-t* som et derivationssuffiks, da det kan følges af en (kompositionel) rod (jf. *alvorligtudseende*), mens et fleksionssuffiks kun kan følges af et andet fleksionssuffiks eller af ordafslutning.

Også talordenes suffigering er morfotaktisk derivativ snarere end fleksiv, idet deres suffikser indgår i komposita (jf. *syv|ende•plads* og *syv|endel•dels•arving*). Uregelmæssige ord for ordenstal og brøker (f.eks. *første*, *anden*, *tredje* og *halv*) markeres individuelt i leksikon. Talordenes form er af betydning for den syntaktiske analyse og tagges i overensstemmelse hermed.

Som præfikser betragtes både leksemer som ikke kan forekomme alene (jf. *sam|leksistens* og *u|ren*) og leksemer som kan forekomme selvstændigt, men som i sammensætninger synes underordnet en stærkere rod (jf. *sammen|sætte* og *mod|arbejde*).

Derivative suffikser er i reglen ukontroversielle. Med ganske få undtagelser er de klart definerede mht. hvilke ordklasser de kan følge og hvilke ordklasser de danner, og de kan derfor tagges med ret stor éntydighed.

3.3 Komposition

Dannelsen af komposita foregår efter følgende mønster:

ORD 1	evt. fuge-s/e	ORD2
subst.	<i>e</i> eller <i>-s</i>	subst., adj. og verbum
adj.	<i>e</i>	subst., adj. og verbum
verbum	<i>ø</i>	subst. og adj.

Med undtagelse af ganske få ord kan ikke-terminale led i et kompositum kun være i grundform, dvs. den morfologisk umarkerede form for substantiver og adjektiver,

og infinitiv for verberne.⁹⁾ De få ord, der kan danne kompositum efter en pluralisform er ført op i et eget leksikon. Produktiv er vist kun *børn*, selv om Diderichsen blandt andre også nævner *bønder*, *fædre* og *mødre* (jfr. f.eks. *mødrehjælpen*).

4 Allomorfi

Helsinki-modellens morfologiske beskrivelse var oprindelig møntet på finsk; redskabet til at beskrive allomorforhold kan derfor synes overdimensioneret i forhold til hvad dansk har brug for, men det er ikke des mindre udmærket anvendeligt.

Gennem hele analysen arbejdes der på to niveauer. Et overfladeniveau, som repræsenterer et ords faktiske tekstform. Og et leksikalt niveau, som repræsenterer den form ordet har i leksikon.

I dansk er disse to niveauer ofte ens, men følgende tre tilfælde retfærdiggør anvendelsen af to-niveau-modellen.

1) **Konsonantfordobling** findes som a) absolut, b) optionel eller c) selektiv.

a) Den absolutte finder altid sted når konsonanten følges af et *-e*, jf. *tak—takken—takker*.

b) Den optionelle kan fordobles ved samme vilkår som a). Det hedder en *virus*, men enten *virussen* eller *virusen* i bestemt form, og enten *viruser* eller *virusser* i pluralis (eller *virus* eller *vira*, men det er konsonantfordoblingen uvedkommende).

c) Den selektive fordobler kun ved efterfølgende pluralismorfem på *-e*, jf. *mad—maden—madder*.

2) Der er fem forskellige allomorforiske varianter af **e-synkope** ved suffigering af substantiver der ender på *-el*, *-en* eller *-er*.

<u>grundform</u>	<u>best. form</u>	<u>pluralis</u>
------------------	-------------------	-----------------

9) Diderichsen nævner enkelte verbum-verbum komposita hvor der bruges stamme-form af verbet, f.eks. *snorksove*, men de er meget sjældne, og samtlige hans eksempler er ført ind i leksikon som selvstændige lemmata.

a. hilsen	hils(e)n en	hils(e)n er
b. neger	neger en	neg(e)r e
c. finger	finger en	fingr e
d. cykel	cyk(e) en	cykl er
e. mirakel	mirakl et	mirakl er

Parenteserne angiver hvor *-e* er optionelt. Herudover er der behov for angivelse af synkope ved fleksionen af et stort antal substantiver der ender på tryksvagt *-e* (jf. *kvinden*, *aber*, modsat *veen*, *patéer*.)

3) Der findes fire forskellige typer **omlyd** ved pluralisformerne af substantiverne¹⁰⁾, jf.

a : æ mand : mænd

a : ø barn : børn

o : ø ko : køer

å : æ tå : tæer

Her er anført overfladeformer. De specielle leksikale former, der giver mulighed for at beskrive de tilladte allomorfer, er invariable; således indeholder f.eks. ordene *tæer* og *tåen* begge leksemet *tå* i de leksikale former (jf. *tå* 'ler og *tå* 'len — " står for omlyd foran pluralis-suffiks). På samme måde repræsenteres ordene *negeren* og *negre* ved de leksikale former *negΣr|en* og *negΣr|e* ('Σ' står for mulig realisation som *-e* under specifikke kontekstuelle vilkår).

Analysen udelukker former som **koer*, **tæer*, **mandene* og **barnene*, mens det gennem et samspil mellem to-niveau-regler og morfotaktisk beskrivelse er muligt at hæfte trækket 'pluralis' på *mænd*, og trækket 'singularis' på *mand*, selvom der ikke er en morfologisk markering af pluralis.

Ovenstående er en gennemgang af de træk af dansk vi har valgt at beskrive dels gennem kontinuationer og det morfologiske system og dels som allomorfer. Resten, som er en relativt lille, men meget frekvent, gruppe af ord, er skrevet ind som opslagsord i leksikon.

Dette gælder for substantiver som *verbum—verbet—verber*, og for mange

10) Der findes som bekendt også omlyd i andre fleksionsparadigmer, men den er for uregelmæssig til at systemet med fordel kan regelsætte den, og behandles derfor individuelt.

andre af samme type med en latinsk eller fremmedartet pluralis.

Det gælder for verber der skifter rod midt i et paradigme, som *gå* med formen *gik* i præteritum, og det gælder ord som *drikke* med formen *drak* i præteritum og formen *drukke* i præteritum participium.

Også for adjektiverne gælder det ord med rodskifte, f.eks. ordet *stor* der kompareres *større* og *størst* og ligeledes *lille*, *mindre*, *mindst*.

5. Resultatet af den morfologiske analyse

Denne implementation af dansk morfologi er i stand til at foretage 'morfologiske dekompositioner' af et uendeligt antal tekstord. En morfologisk dekomposition er en analyse af et ord, hvor ordet dels er opløst i mindstede, dels er grammatisk karakteriseret; desuden er ordets lemmatiske grundform anført. En grammatisk karakteristisk indeholder forskellige træk for de forskellige ordtyper.

Nominalerne (substantiver, pronominer og adjektiver) er karakteriseret mht. ordklasse, genus, bestemthed, numerus og kasus. Dvs. at et nominal vil blive taget med en kombination af følgende:

Klasse	Genus	Bestemthed	Numerus	Kasus
N	FÆL	UBEST	SG	NOM
PRON	INT	BEST	PL	(AKK)
A	FÆL/INT	UBEST/BEST	SG/PL	GEN
	?	?	?	?

Dobbeltangivelser (som 'fæl/int') er naturligvis mindre nyttige for den videre analyse; men de angiver dog at det pågældende træk *kan* være relevant. F.eks. vil i sekvensen *et ringe resultat* ordet *ringe* være taget 'fæl/int', til angivelse af at ordet accepterer kongruens med *resultat*, uanset det ikke er morfologisk markeret. Betegnelsen '?' angiver at trækket er irrelevant, som f.eks. bestemthed ved egennavne. Akkusativ er kun muligt ved de personlige pronominer. Adjektiverne får desuden en markering for komparation.

For **verberne** er det implicit angivet om de er finitte (imperativ, præsens eller præteritum) eller infinitte (infinitiv eller participium). Et verbum vil iøvrigt blive taget med en kombination af følgende:

Klasse	Passivitet	Finithed
V	AKT	PRÆS
	PASS	PRÆT
		IMP
		INF
		PCP1 (præs.part.) + kasus-markering
		PCP2 (præt.part.) + genus, bestemthed, numerus og kasus-markering

Udover dette er der fra leksikon sat en (syntaktisk) ordklasse på resten af ordene (i henhold til RO), hvoraf talord, sideordnende konjunktioner, forkortelser og udråbsord i reglen er éntydige, mens betegnelserne adverbium, underordnende konjunktion og præposition sjældent forekommer alene.

Det er vigtigt at mærke sig forskellen mellem morfologisk definerede tags (de primære), og syntaktisk definerede tags (de sekundære). Sidstnævnte repræsenterer en hypotese mht. ordenes syntaktiske funktion, som kun den videre syntaktiske analyse kan be- eller afkræfte. Hypotesen er berettiget, idet den hjælper til at bestemme 999 ud af 1000 forekomster af f.eks. ordet *men*. Men — og det er et vigtigt men — der kan også forekomme andre anvendelser af ordet.

Resultatet af den automatiske tagging er så en tekst hvor alle ord, for så vidt de kan findes i leksikon, har fået én eller flere analyser¹¹⁾, hver med et antal tags. Et sådant sæt af analyser betegnes en 'kohorte'. Det er altafgørende at den korrekte analyse forefindes i kohorten. Det er til gengæld i første omgang mindre væsentligt hvor mange kohorter der kun indeholder én analyse (selv om én analyse pr. ord naturligvis er målet for disambigueringen); for der skal under alle omstændigheder foregå en videre bearbejdning, med henblik på disambiguering, lemmatisering og syntaktisk analyse — disse ting går hånd i hånd.

6 Den videre analyse

Det er ikke her stedet at redegøre i detaljer for den videre analyse — den

11) Ord som systemet ikke kan analysere, markeres specielt; og det er naturligvis en væsentlig opgave løbende at undersøge listen over ikke-analyserede ord, med henblik på nødvendigheden af en berigelse af leksikon eller af den grammatiske beskrivelse.

foreligger for vores vedkommende kun som plan og principper (jf. Arndt in press). En langt mere fyldestgørende beskrivelse af det arbejde der allerede er gjort med engelsk og svensk kan findes i Karlsson et al. 1991 og Karlsson 1992. Den følgende skitse er medtaget for at give et indtryk af arbejdet i sin helhed.

6.1 Den dekontekstuelle disambiguering

Analysen er behæftet med en del kompositionelle overgenereringer af typen *kul•tur* og *kult•ur* fra indledningen, analyser som hverken dækker over en guidet tur til en mineby eller et Rolex-ur. Det skyldes at komposition og derivation er overordentlig produktive fænomener på dansk, og at analysen derfor kan acceptere ikke alene de frekvente *borende*, *kultur* og *solsikker*, men også de mindre sandsynlige (men dog mulige) *bor•ende*, *kult•ur* og *sol•sikker* (sidstnævnte f.eks. om en creme der beskytter mod ultraviolette stråler).

Erfaringer fra analysen af svensk (jf. Karlsson 1992) har vist at et meget stort antal af sådanne fejl kan elimineres ved to simple regler: I en kohorte af flere analyser foretrækkes den (eller de) der har færrest kompositions-grænser, og dernæst den (eller de) der har færrest derivations-grænser. Det er anvendelsen af disse regler der forudsætter at almindeligt forekommende derivationer og kompositioner findes som simplex-ord i leksikon (jf. ovenfor note 3 og sektion 2). Ellers ville reglerne altid foretrække *ur|et* fremfor *u|ret* og *bord|ende* fremfor *bord•ende*, osv., hvilket naturligvis ikke er acceptabelt. Til gengæld skaffer reglerne os ubesværet af med misdannelser som *mod|el* og *mis|te*, til fordel for *model* og *miste*.

Udvidelsen af leksikon med komplekse ord er i god overensstemmelse med vores sprogfornemmelse: de pågældende ord føles altovervejende som simple dele af det leksikale inventar, ikke som produktive dannelser (uanset de kan analyseres morfologisk). Men definitionen af 'almindeligt forekommende' er naturligvis flydende.

6.2 Kontekstuel disambiguering og syntaktisk analyse

Efter den dekontekstuelle disambiguering forventer vi at ca. 60% af ordene er morfologisk éntyde, og at der er meget få ord med mere end 4 forskellige

analyser (jf. Karlsson 1992; foreløbige analyser viser at disse tal er sandsynlige også for dansk).

En simpel kontekstuel disambiguering kan ske ved regler, der er baseret på de allerede angivne typer tags, og som kan formalisere f.eks. at "vi ikke har to på hinanden følgende finitte verber på dansk". Det lyder ret simpelt, men reglen skal tilføjes et forbehold: "i en ikke-kompleks sætning". Og det komplicerer beskrivelsen betragteligt.

Forudsætningen for en tilstrækkelig kompleks syntaktisk analyse er, at der til hvert ords tag føjes yderligere syntaktiske træk, ud over de ovenfor nævnte simple ordklassebetegnelser. Disse tags baseres på topologi og sætningssyntaks, samt verbal valens, men de kan tilføjes automatisk på grundlag af den allerede bestemte leksikale og/eller morfologiske karakteristik, for så vidt man stadig fastholder en rigelig overbestemmelse af de syntaktiske muligheder. Et simplificeret eksempel vil vise princippet for frasortering af syntaktisk uacceptable tags (der er kun medtaget et udpluk af de tag-elementer der kan forefindes):

1	2	3	4	5	6	7
De	blinde	skal	også	kunne	komme	ind.
NOMPL*	ATTRIB*	NOMSG*	ADV	PRÆT*	INFIN	ADV
AREPL	PRÆD*	PRÆS		INFIN	NOMSG*	
	NOMPL					
	NOMSG*					

I kolonne 2 frasorteres PRÆD, fordi et prædikativt adjektiv ikke kan forekomme hverken mellem artikel og nominal eller mellem pronomen og finit verbum. ATTRIB frasorteres fordi artiklen på den ene side er pluralis og nominet på den anden er singularis. Af samme grund frasorteres NOMSG i både kolonne 2 og 3. Dernæst frasorteres i kolonne 5 PRÆT, fordi der allerede er et finit verbum i denne simplex-sætning. Endelig frasorteres i kolonne 6 NOMSG, fordi modalverbet *kunne* ikke kan tage et adverbialt komplement (slet ikke ved siden af et substantivisk)¹²⁾. Brugen af ordet 'frasortere' her er signifikant: Karlsson (1990)

12) For en ordens skyld bemærkes at disse regler faktisk ikke er tilstrækkelig komplekse. Men de er tilstrækkelige til at vise princippet.

betegner af samme grund den syntaktiske komponent som en 'constraint grammar' — den 'udgrænser' de alternativer der er syntaktisk uacceptable.

7 Afsluttende bemærkninger

Modellen kan ikke disambiguere det man kunne kalde 'ægte tvetydigheder'. En sætning som *Når jeg er deprimeret, går jeg hen og køber medicin og piller ved den søde apotekerfrue* vil få to analyser. Det skal den også have.

Modellen er i princippet både produktiv og receptiv: reglerne skulle kunne både generere og analysere tekster. I praksis har vi holdt os til den analytiske side; f.eks. kan vores regler for komposition ikke forhindre overgenerering af både besynderlige komposita og forkert anbragte 'fuger'; også reglen om adjektivers komparation overgenererer, osv.

Men uanset de begrænsninger der kan findes i modellen, så muliggør den en lemmatisering og en syntaktisk tagging af korpora. Det finder vi er et spændende perspektiv for sprogforskningen.

Referencer

- Antworth, Evan L. (1990) *PC-KIMMO, A Two Level Processor for Morphological Analysis* Dallas, Texas
- Arndt, Hans (in press) "Towards a syntactic analysis of Danish computer corpora" in *Proceedings of the 13th Scandinavian Conference of Linguistics*, Roskilde Universitetscenter 1992
- Bergenholtz, Henning (1991) "DK87-DK90: Dansk korpus med almensproglige tekster" in Mette Kunø og Erik Vive Larsen 3. *Møde om Udforskningen af Dansk Sprog* Aarhus Universitet
- Bilgram, Thomas (1992) "Et maskinlæsbart leksikon" Eksamensopgave, Institut for Lingvistik, Aarhus Universitet
- Garside, Roger, Geoffrey Leech & Geoffrey Sampson (eds) (1987) *The Computational Analysis of English: A Corpus-Based Approach* London: Longman
- Hansen, Peter Molbæk (1989) "Nogle svagheder ved toniveaumorfologien" in *SAML*, Institut for Almen og Anvendt Lingvistik, Københavns Universitet
- Karlisson, Fred (1990) "Constraint Grammar as a Framework for Parsing Running Text" in Hans Karlgren (ed.) *Proceedings from the XIIIth International Conference on Computational Linguistics* Helsinki
- Karlisson, Fred (1992) "SWETWOL: A Comprehensive Morphological Analyzer for Swedish" in *Nordic Journal of Linguistics*
- Karlisson, Fred, Atro Voutilainen, Arto Anttila & Juha Heikkilä (1991) "Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text, with an Application to English" in *Workshop Notes from the Ninth National*

Conference on Artificial Intelligence by the American Association for Artificial Intelligence, Anaheim, California, 1991

Retskrivningsordbogen (1986) Dansk Sprognævn, Gyldendal (disketteversion, venligst stillet til rådighed af redaktionen)

Sinclair, John (1991) *Corpus, Concordance, Collocation* Oxford University Press