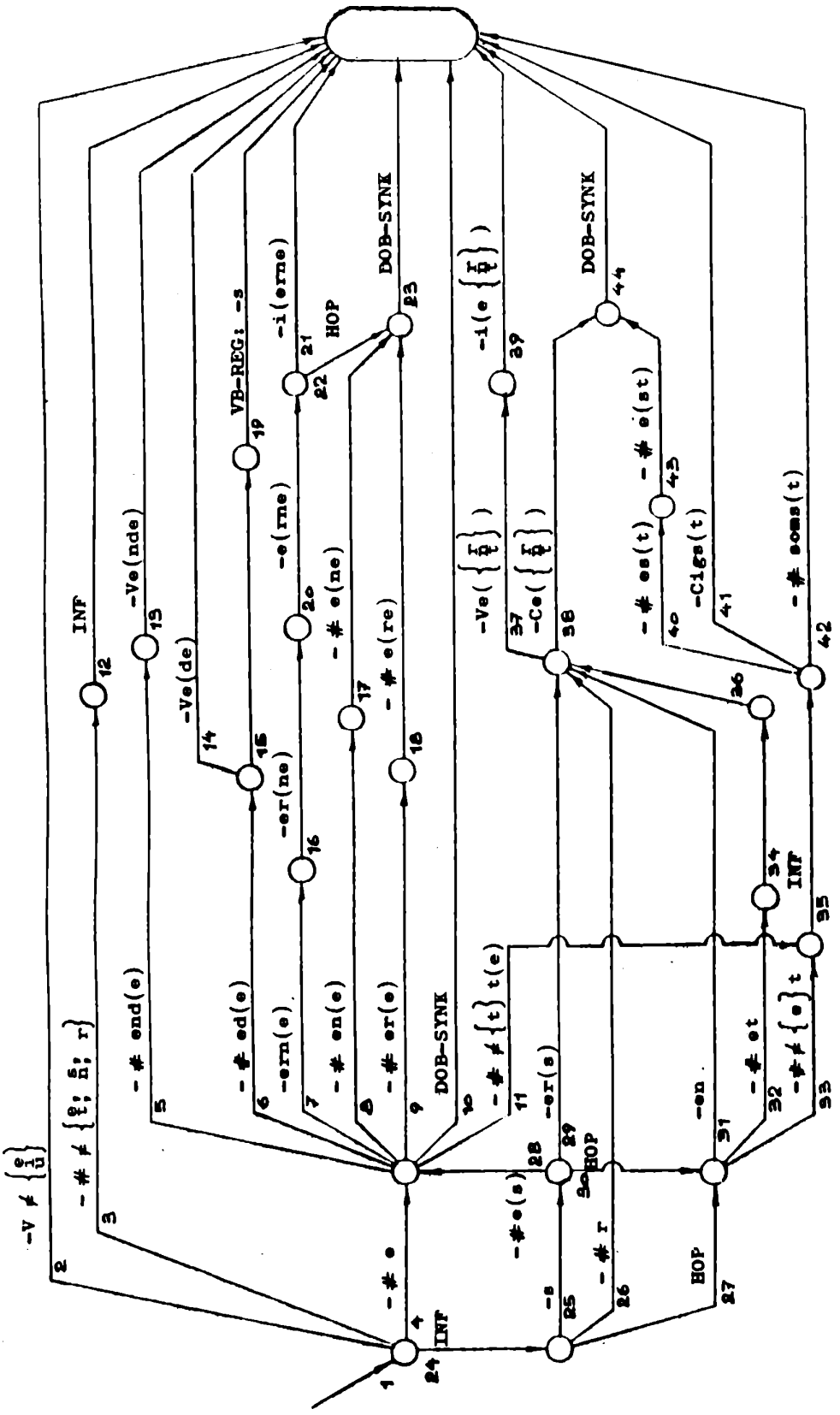
Hanne Ruus
Institut for nordisk filologi
Københavns Universitet
Njalsgade 80
DK-2300 København S.

DANSK MORFOLOGI TIL AUTOMATISK ANALYSE.

Den datamatiske analyse af naturlige sprog kan have de mest forskelligartede formål. Man kan have et sprogvidenskabeligt formål: Ved at formulere dele af et naturligt sprogs beskrivelse i grammatikker beregnet til at anvendes af en parser (jf. andre bidrag i denne publikation) og udvælge passende varierede data, kan man kontrollere, om den lingvistiske beskrivelse er udtømmende og sammenhængende. Man kan have et matematisk formål: Man kan afprøve, hvilken type formelt system og dertil svarende automat der egner sig til at beskrive en given del af et naturligt sprogs grammatik. Endvidere kan man have et formål, der er defineret af en bestemt anvendelse af den datamatiske analyse af naturlige sprog.

Når man udarbejder et stykke sprogbeskrivelse, vil anvendelsen af denne beskrivelse være medbestemmende for, hvordan beskrivelsen vil komme til at se ud; i øvrigt vil beskrivelsen selvfølgelig afhænge af det sprog og det fænomen, man beskriver. De nordiske sprog har en relativt kompliceret suffigerende fleksion, sammenlignet med andre vesteuropæiske sprog som engelsk, tysk og fransk. Alle fleksionskategorier kan udtrykkes med suffikser. Det betyder, at de store fleksionsordklasser substantiver, verber og adjektiver alle har ordformer med mere end ét suffigeret fleksiv. En maksimal substantivform er f.eks. pigernes, som er pige bøjet i numerus (pluralis: -r), i bekendthed (-ne) og i kasus (genitiv: -s). En maksimal verbalform er ventedes, der er bøjet i tempus (præteritum: -de) og i diatese (passiv: -s). En maksimal adjektivform vil være den yngstes bøjet i komparation (superlativ: -ste) og kasus (genitiv: -s). Før et eller flere fleksiver kan et ords stamme kræve forskellige morfografematiske ændringer som konsonantfordobling eller -forenkling og/eller synkope. Formen gamle er f.eks. stammen gammel ændret ved tilføjelse af fleksivet -e, synkope (tab af e'et foran l) og konsonantforenkling.

Dansk morfologi til automatisk analyse
Hanne Ruus
Proceedings of NODALIDA 1983, pages 178-186



Figur 1. (fra Ruus 1977).

Den nordiske fleksion er også karakteristisk ved sit ringe inventar af endelser, som så til gengæld hver især har flere forskellige betydninger. I dansk bruges f.eks. -r både af substantiver og verber, -e af både substantiver, verber og adjektiver. I norrønt sprog bruges -a, -ar, -ir i både substantiv-, verbal- og adjektivbøjningen.

Figur 1. viser en udtømmende beskrivelse af den regelmæssige danske substantiv-, verbal- og adjektivbøjning. Man går ind i netværket til venstre. De bøjede ord undersøges bagfra. Når en del af en endelse passer med betingelsen på en pil, fjernes denne del og man fortsætter i den knude, som pilen fører til. DOB-SYNK refererer til et delnetværk, hvor de morfografematiske ændringer behandles (jf. Ruus 1977, Ruus 1978). Denne grammatik giver en udtømmende beskrivelse af de mulige endelser, deres indbyrdes kombinationsmuligheder og deres optræden sammen med morfografematiske ændringer. Der er ikke i denne grammatik taget stilling til, hvilken eller hvilke fortolkninger de enkelte endelser og endelseskombinationer skal have. Når man bruger grammatikken over dansk fleksion til at vise, at den er en udtømmende sproglig beskrivelse, vil man udfolde den enkelte endelses potentiale i overensstemmelse med almindelig sprogvidenskabelig praksis. I adjektivbøjningen vil -e f.eks. blive beskrevet som enten bestemt form singularis eller ubestemt form pluralis eller bestemt form pluralis. Hvis man skal bruge de morfologiske oplysninger i en praktisk anvendelse, kan det være hensigtsmæssigt at udskyde fortolkningen af en endelse til et senere trin i analysen.

Blandt de mange områder, hvor man arbejder med datamatisk analyse af naturlige sprog til bestemte praktiske formål, er maskinoversættelse et af de mest ambitiøse (Ruus 1984). I maskinoversættelsessystemer beregner man en syntaktisk-semantisk struktur for sætningerne i kildeteksten og bruger denne strukturs informationer til at finde de rigtige målsprogsækvivalenter (se f.eks. om Eurotra Maegaard og Ruus 1980, Ruus og Maegaard 1982). Hvordan man beregner den syntaktisk-semantiske struktur, afhænger af, hvilken lingvistisk strategi man vælger. For at kunne beregne den må man dog have nogle oplysninger om de sekvenser af tegn, som inddata består af. Her kan man inddrage den morfologiske analyse.

Man kunne spørge, om man ikke kan klare sig med en fuldformsordbog. I en sådan vil alle bøjede ordformer forekomme som selvstæn-

dige opslag ledsaget af de relevante oplysninger om bøjningsform og betydning. Et maskinoversættelsessystem er et generelt system. Det skal kunne genkende og behandle alle ord i det analyserede sprog. Til denne anvendelse vil det være umuligt for sprog som de nordiske at fremstille en komplet fuldformsordbog, da nye sammensatte og afledte ord dannes løbende.Selv med en større fuldformsordbog måtte man sørge for at have et system til at behandle ukendte ordformer bl.a. ved at fjerne bøjningsendelser. Når man skal have et sådant system, kan man lige så godt bruge det til alle ordformer. Så kan man nøjes med at lagre lemmaer i ordbogen. Hvert lemma skal så forsynes med oplysninger om bøjning, syntaktisk klasse og betydninger.

Man skal altså bruge en fleksionsanalyse. Hvilke oplysninger kan fleksionsanalysen give? Fleksiver kan levere to slags oplysninger: De fleksiver, der har en betydning, repræsenterer den del af lingvistisk semantik, der er bedst udforsket. Det gælder f.eks. numerus- og tempusfleksiver. De fleksiver, der ikke har betydning, kan ofte bruges som vejvisere i den syntaktiske analyse. Det gælder f.eks. genusfleksiver.

Blandt andet som følge af de multianvendelige endelser er de nordiske ord ofte flertydige. Med Allén (Allén 1970 p. XIX) kan man skelne mellem ekstern homografi, dvs. at en ordform kan tilhøre forskellige syntaktiske klasser f.eks. murer sb. og murer vb. i præsens, og intern homografi, dvs. at en ordform kan være forskellige former af samme lemma f.eks. ord, som kan være singularis eller pluralis. Selv ved en ganske enkel fleksivanalyse med ordbogsopslag vil man altså ofte få to og flere løsningsforslag til de enkelte ordformer i den tekst, der skal analyseres.

I det følgende skal jeg skitsere en metode til at behandle interne homografer, sådan at de får én og kun én beskrivelse af fleksionsanalysen og dernæst fortolkes i den syntaktiske analyse. Som eksempel vælges den danske adjektivbøjning i positiv i kombination med neutrale substantiver.

Ud fra bøjningen i positiv kan man udskille fire slags danske adjektiver:

	endelser	eksempel
type I	-Ø -t -e	sød
type II	-Ø -t	grå
type III	-Ø -e	sort frisk
type IV	-Ø	ringe

Som man ser, optræder Ø-endelsen ved alle fire typer adjektiver. Det giver i alle tilfælde anledning til intern homografi, men af forskellig art:

artikel	type I	type II	type III	type IV	substantiv
	sød-Ø	grå-Ø	sort-Ø	ringe-Ø	(fælleskøn)
	sød-t	grå-t	sort-Ø	ringe-Ø	øl, brød
et	sød-t	grå-t	sort-Ø	ringe-Ø	brød
det	sød-e	grå-Ø	sort-e	ringe-Ø	brød
	sød-e	grå-Ø	sort-e	ringe-Ø	brød
de	sød-e	grå-Ø	sort-e	ringe-Ø	brød

Man ser af ovenstående oversigt, at Ø-endelsen bruges i alle kontekster ved type-IV-adjektiver, ved type-II og type-III bruges Ø-endelsen, henholdsvis hvor type-I har Ø- og e-endelse og hvor type-I har Ø- og t-endelse. Det er derfor nødvendigt at operere med 4 forskellige Ø-endelser ved adjektiver i positiv. Hvis vi nummererer dem efter typerne ovenfor, kan de fire adjektiver have følgende oplysninger i ordbogen om deres bøjning i positiv:

sød	adj-Ø1, adj-t, adj-e
grå	adj-Ø2, adj-t
sort	adj-Ø3, adj-e
ringe	adj-Ø4

For de forskellige former af de fire adjektiver leverer fleksionsanalysen følgende oplysninger:

	inddata	analyseresultat
type I	sød*	sød+adj-Ø1
	søde*	sød+adj-e
	sødt	sød+adj-t

type II	grå	grå+adj-Ø2
	gråt	grå+adj-t
type III	sort*	sort+adj-Ø3
	sorte	sort+adj-e
type IV	ringe*	ringe+adj-Ø4

En analyse af pluralisbøjningen af neutrale substantiver vil vise, at man her har 2 forskellige Ø-endelser:

artikel	type I	type II	type III
det	brød-Ø	hus-Ø	æble-Ø
de	brød-Ø	hus-e	æble-r

Alle tre typer substantiver har Ø-endelse i singularis, type-I har også Ø-endelse i pluralis, hvor de andre typer har e- og r-endelse. En del af fleksionsoplysningerne ved disse ord i ordbogen kunne da se således ud:

brød	sbneu-Ø1
hus	sbneu-Ø2, sb-e
æble	sbneu-Ø2, sb-r

Fra fleksionsanalyse med ordbogsopslag får vi da følgende resultater for former af disse ord:

	inddata	analyseresultat
type I	brød*	brød+sbneu-Ø1
type II	hus*	hus+sbneu-Ø2
	huse*	hus+sb-e
type III	æble	æble+sbneu-Ø2
	æbler	æble+sb-r

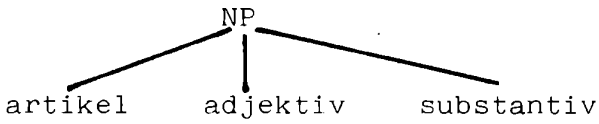
Man ser, at den enkelte ordform kun får et analyseresultat med den omhandlede ordklasse. * ved ordformerne angiver, at de har ekstern homografi og derfor også vil blive analyseret som former af andre ordklasser.

Det står nu tilbage at skitsere, hvordan den syntaktiske analyse kan fortolke fleksivoplysningerne, når den danner nominalhypotagmer. Vi antager, at den syntaktiske analyse bl.a. indeholder en grammatik, der bygger nominalhypotagmer, og at den er opdelt i undergrammatikker, der bygger forskellige slags nominalhypotagmer. Én undergrammatik bygger hypotagmer med neutrale ord i singularis som kerne

(1) en regel med betingelsen

$$\underline{et} + \text{adjektiv} + \begin{Bmatrix} \text{adj-t} \\ \text{adj-Ø3} \\ \text{adj-Ø4} \end{Bmatrix} + \text{substantiv} + \begin{Bmatrix} \text{sbneu-Ø1} \\ \text{sbneu-Ø2} \end{Bmatrix}$$

vil bygge et hypotagme:



I reglen vil man så også sørge for, at egenskaberne ubestemt og singularis bliver knyttet til hypotagmet.

(2) en regel med betingelsen

$$\underline{\text{det}} + \text{adjektiv} + \left\{\begin{array}{l}\text{adj-e}\\ \text{adj-Ø2}\\ \text{adj-Ø4}\end{array}\right\} + \text{substantiv} + \left\{\begin{array}{l}\text{sbneu-Ø1}\\ \text{sbneu-Ø2}\end{array}\right\}$$

vil bygge et hypotagme med samme struktur som regel (1), men her vil egenskaberne bestemt og singularis blive knyttet til hypotagmet.

Hvis man sammenholder de syntaktiske reglers betingelser med fleksionsanalyseresultaterne ovenfor, vil man se, at analysen af adjektivformerne sødt, gråt, sort, ringe passer med betingelsen for adjektivet i regel (1) og analysen af formerne søde, grå, sorte, ringe med betingelsen i regel (2). De flertydige former søde, grå, sorte, ringe får således en rigtig tolkning, uden at det har været nødvendigt at søge mellem alternative tolkningsforslag. Det samme gælder for den flertydige substantivform brød, der her bliver tolket som singularis, fordi den står i en singulariskontekst.

Da genus ikke markeres i pluralis på dansk, vil reglerne for konstruktion af nominalhypotagmer i pluralis skulle behandle både neutrale substantiver og fælleskønssubstantiver. Grammatikregler for hypotagmer i pluralis kan derfor først formuleres, når bøjningen af fælleskønssubstantiver er analyseret for flertydige endelser.

Den her skitserede behandling af interne homografer skulle da have følgende fordele: der gives kun ét fleksionsanalyseresultat for de forskellige tolkningsmuligheder af en intern homograf. Den rigtige tolkning vælges i den syntaktiske analyse, hvor man bygger på konteksten. I de sjældne tilfælde, hvor den nærmeste kontekst er flertydig på samme måde for både substantiv

og adjektiv f.eks. ringe brød med analysen ringe+adj-Ø4+brød+ sbneu-Ø1, vil analysen selvfølgelig passe både til betingelsen på en regel, der danner nøgne nominalhypotagmer i singularis, og på én, der danner nøgne nominalhypotagmer i pluralis. Hvis man i en given tekst skal vælge den rigtige af disse to løsninger, må man inddrage en større kontekst.

Ved behandling af store tekstmængder vil det være en tidsmæssig fordel, at interne homografer kun får én analyse i fleksionsanalysen, der skal gennemløbes for alle sætninger. De omtalte interne homografier hører nemlig ligesom de interne homografier i verbalbøjningen til blandt de almindeligste bøjningsformer (Noesgaard 1960).

Den beskrevne strategi kan muligvis også anvendes ved ekstern homografi som den forekommer ved mange funktionsord f.eks. den, det, de artikler eller pronominer. Hvis disse ords hele betydning fremgår af deres syntaktiske placering, kan de optræde eksplicit i de syntaktiske regler som eksemplificeret ovenfor og vil derfor ikke kræve ordbogsoplysninger og morfologisk analyse. I et maskinoversættelsessystem, der jo skal analysere alle slags tekster og mange af dem, vil det være en tidsmæssig gevinst, hvis sådanne ord kan nøjes med en meget simpel eller ingen indledende analyse. Disse ord hører til blandt de allerhyppigste ord, hvor en beregning har vist, at de 40 hyppigste homografer udgør 31% af en tekst (Maegaard og Ruus 1980b).Om den betydningsbeskrivelse, der kan deduceres i analysen, er tilstrækkelig for disse ord til maskinel oversættelse, vil vise sig, når man kommer til praktiske eksperimenter med at producere oversættelser til og fra dansk, forhåbentlig i løbet af det næste par år.

Litteraturhenvisninger.

Allén, Sture 1970: Nusvensk Frekvensordbok bd.I
Maegaard, Bente and Ruus, Hanne 1980: Structuring Linguistic Information for Machine Translation, The EUROTRA Interface Structure, i Human Translation, Machine Translation ed. Suzanne Hanon and Viggo Hjørnager Pedersen = NOK 39, Romansk Institut, Odense Universitet, p. 159-169.
Maegaard, Bente og Hanne Ruus: Danske almindelige ord. Rangfrekvenslister og deres brug, i SAML 7, p. 5-22.
Noesgaard, A. 1960: Hyppighedsundersøgelser III.

Ruus, Hanne 1977: Ordmekanik, i SAML III, p.79-106.
Ruus, Hanne 1978: Suffix-stripping. Automatisk lemmatisering af regelmæssigt bøjede danske ord, i Selskab for nordisk filologi, Årsberetning for 1974-76, p.14-21.
Ruus, Hanne and Maegaard, Bente: Multilingual Syntax and Morphology for Machine Translation, i Machine Translation and Computational Lexicography,ed. Karl Hyldgaard-Jensen and Bente Maegaard, p. 26-34.
Ruus, Hanne 1984: Investigating the Nordic Languages for the Information Society, udkommer i Proceedings from V. International Conference of Nordic Languages and Modern Linguistics, Århus.

Øvrige bidrag om parsing i nærværende publikation.

Forkortelse:
SAML = Skrifter om Anvendt og Matematisk Lingvistik, udgivet af Institut for anvendt og matematisk lingvistik, Københavns Universitet.