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



# SEMANTIK I AUTOMATISK LEMMATISERING.

De fleste, der beskæftiger sig med lemmatisering, har som data autentiske tekster på naturlige sprog og som nært mål at kunne svare på spørgsmål som: Hvor mange og hvilke ord er almindelige i aviser, i en forfatters værker, i børnebøger osv. ?

I denne situation er der flere grunde til at automatisere lemmatiseringsprocessen. For det første den rent praktiske, at en så godt som automatisk lemmatisering gør det muligt at komme igennem et stort materiale på overskuelig tid. For det andet den vigtigere grund, at kun ved at beskrive tilordningen af bøjningsformer til lemmaer algoritmisk kan man sikre sig en helt konsekvent behandling af materialet. For det tredie kan en automatiseret lemmatiseringsprocedure anvendes igen og igen på mange slags tekst til forskellige formål: frekvensundersøgelser, automatisk syntaktisk analyse, maskinoversættelse.

En automatisering fordrer en entydig definition af lemma: man må kræve udtryksforskel i mindst én form i to serier af ordformer for at opstille to lemmaer (jf. Nusvensk Frekvensordboks og DANwORDs[1] lemmadefinition). Selv med denne formelle lemmadefinition kommer man ikke uden om at anvende semantiske oplysninger, når man vil nærme sig en fuldautomatisk lemmatisering. For at få en enkel og billig lemmatiseringsprocedure må man imidlertid vride så meget information som muligt ud af udtryksforskelle og begrænse sig til et minimum af semantik. Derfor er jeg på udkig efter semantiske træk, der kan bruges ved mange forskellige ord, og helst sådanne, som kobles med udtryksforskelle.

Heterografer er ord, der i alle bøjningsformer staves anderledes end alle andre ord. De kan altså lemmatiseres automatisk alene ud fra deres udseende - uden semantik. Men vejen til den automatiske analyse af naturlige sprog er brolagt med entydiggjorte homografer. Man kan komme et godt stykke i entydiggørelsen ved hjælp af formelle træk i konteksten, men derefter er man henvist til semantikken.

Overvejelserne om entydiggørelse nedenfor bygger på excerpering af ordformerne i fig. 1 og 2 i DANwORDs[1] prøver fra fiktionstekster for voksne (godt 250 000 løbende ord), hvad der gav godt 200 belæg på rejse I, II og III og godt 50 belæg på øre I og II. Eksemplerne er valgt, så de viser homografi inden for samme ordklasse: øre I og II er substantiver, rejse II og III er verber.

Semantik i automatisk lemmatisering
Hanne Ruus
Proceedings of NODALIDA 1979, pages 139-143

øre I "legemsdel" sb. -t, pl. -r el. -n

| | |
|---|---|
| øre | |
| øret | |
| ører | |
| (øren) | |
| ørerne | |
| ørene | |

øre II "betalingsmiddel" sb. -n, pl. dsm. el -r

| | |
|---|---|
| øre | |
| (øren) | |
| ører | |
| ørerne | |
| ørene | |

fig. 1.

I figuren er entydige former streget under. Da de to substantiver har hvert sit genus, kan en del forekomster af øre i singularis entydiggøres ud fra kongruensbøjede former foran ordet: dit øre, det indre øre til øre I og en øre til øre II. øre brugt som pluralis af øre II kan også bestemmes ud fra konteksten: det har altid et talord som adled tolv øre, halvtreds øre. Men ved de øvrige homografe pluralisformer er man henvist til semantiske hjælpemidler.

rejse I sb. pl. -r

| |
|---|
| rejse |
| rejser |

| rejse II "travel" vb. -te | rejse III "raise" vb. -te, -ning |
|---|---|
| rejse | rejse |
| rejser | rejser |
| rejste | rejste |
| rejst | rejst |

fig. 2.

---

I figuren er fra substantivet rejse kun anført de former, der er homografe med verbalformer. Disse substantivformer vil i de fleste tilfælde kunne udskilles ved hjælp af kongruerende adled, f.eks. en besværlig rejse, hele denne rejse, den oplevelsesrige rejse. Begge verber er transitive, men kun rejse III kan have refleksivt objekt. Denne syntaktisk-semantiske oplysning er kvantitativt set vigtig, da knap 2/3 af belæggene på rejse er former af rejse sig. Til resten af verbaleksemplerne må man finde mere forfinede, semantiske deskriptorer.

Når man ser sig om i den datalingvistiske forskning for at finde ideer til semantiske oplysninger og anvendelsen af dem, viser det sig, at de projekter, der arbejder med automatisk behandling af semantiske oplysninger, befinder sig inden for området simuleret intelligens, hvor man interesserer sig mere for metoderne end for resultaterne og følgelig udvikler disse metoder på ret begrænsede tekstmængder. Her arbejder man på at afbilde så meget betydning, at maskinen kan simulere en form for forståelse af den indlæste tekst. Man er derfor mindre interesseret i udtrykket og bruger til gengæld komplicerede sæt af semantiske primitiver og slutningsregler, som det i deres nuværende form ville være uoverkommeligt at anvende på større tekstmængder. Men måske kan man låne et lille sæt anvendelige kategorier?

| øre I | øre II | |
|---|---|---|
| HEAD: PART | HEAD: THING | B. HEAD: SIGN |
| QUAL: ANI | QUAL: "metal" | |

fig. 3.

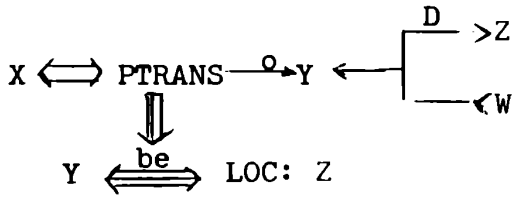
Figuren viser de vigtigste dele af de semantiske udtryk for øre I og II formuleret i Yorick Wilks'[2] semantiske formler. øre II har to betydninger, en der henviser til selve mønten (THING "metal") og en anden til værdien (SIGN), mens øre I er beskrevet som del af menneske eller dyr (PART ANI)

Ved brug af semantiske oplysninger kan man nøjes med at udvælge sådanne, der entydigt kan bestemme forekomster af den ene af to tolkningsmuligheder, hvis man kan finde semantiske oplysninger, der gør dette tilstrækkeligt sikkert.

Da øre II har flere betydninger i Wilks' system, er det oplagt at forsøge at finde træk, der entydigt kan udskille øre I-eksemplerne. Af excerpterne fremgår det, at øre I ofte forekommer med possessiver som adled, jf. de possessive neutrumformer, der blev brugt til at bestemme singularisformerne af øre I. Hvis man tilføjer en regel om, at possessiv ofte står foran "legemsdel", kan man entydiggøre ved hjælp af en delvis formelt afgrænselig klasse af ordformer, de possessive pronominer og personnavne i genitiv, Denne regel vil klare over 3/4 af de resterende flertydige belæg og virker bedre end en ren semantisk regel, der bygger på, at ord for legemsdele ofte forekommer sammen, og derfor forudsætter, at alle ord for legemsdele er mærket som sådanne uanset entydighed. Den semantiske regel ville kun klare halvt så mange eksempler som possessiv-reglen og yderligere kræve, at ansigt, mund, øjne, kæbeben og rottehaler er mærket som legemsdele.

Til entydiggørelse af former af rejse II og III kan man hente hjælp i Roger Schanks Conceptual Analysis[3]. Beskrevet i Schanks diagramform ser rejse II og III således ud:

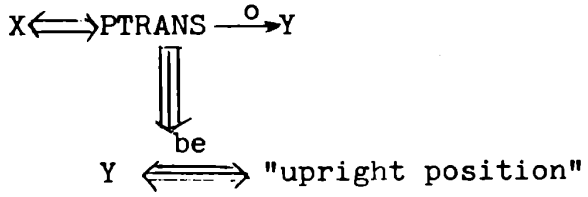
rejse II "travel"




Betingelser til de størrelser, der indgår på pladserne:

X: human; Y = X; Z, W: place

rejse III "raise"




Betingelser til de størrelser, der indgår på pladserne:

X: animate; Y: physobj

fig. 4.

Diagrammet over rejse II kan parafraseres: at X fysisk flytter Y fra W til Z, forårsager, at Y er på lokaliteten Z.

Diagrammet over rejse III kan parafraseres: at X fysisk flytter Y, forårsager, at Y befinder sig i oprejst stilling.

rejse III har altid "ægte" objekt og burde derfor kunne udskilles af den syntaktiske analyse. Denne kan imidlertid ikke laves uden visse semantiske oplysninger: mulige kerner i alle de tids- og målsadverbialer, der kan have form af et nominalhypotagme, må særmærkes. Ellers ville man få forkert analyse af sætninger som Søren rejste en del omkring. Det er altså ikke så lige til at udskille rejse-III-eksemplerne. Man kan derimod bestemme ganske mange af rejse-II-belæggene ved at anvende den semantiske rolle retning, Schanks D-case (Directive). Den manifesteres nemlig, så den er nogenlunde let genkendelig f.eks. ved retningsadverbier og præpositionssyntagmer indledt med til og fra (Hanne Ruus: Sproglig betydningsanalyse, i Nydanske Studier 10-11, 1979, s.186). En regel om at søge efter en retningsangivelse i konteksten vil bestemme godt halvdelen af de resterende flertydige belæg rigtigt.

Entydiggørelsen af den sidste restmængde vil formentlig kræve adgang til ganske fyldige semantiske beskrivelser af mange ord f.eks. kan rejse III have så semantisk forskellige objekter som hoved, sigtelse, spørgsmål, galge og hær.

Denne undersøgelse af 2 sæt homografer har vist, at man forholdsvis let kan opstille syntaktisk-semantiske regler, der ved hjælp af, til dels, formelt afgrænsede klasser, possessiver og retningsadverbier, entydiggør de fleste af de belæg, der ikke kan klares uden semantik.

Som bemærket ovenfor er det vigtigt, at de semantiske træk, man vælger ud til brug i entydiggørelsesprocedurer, kan anvendes ved mange ord og ikke medfører krav om semantiske oplysninger ved alle ord. De træk, jeg har skitseret brugen af her, opfylder begge dette krav: den semantiske rolle "retning", som blev foreslået ved rejse-eksemplerne, vil kunne anvendes ved flertydige bevægelsesverber som fare, føre, lede, mens klassen af possessiver, som blev indført ved øre, vil være nyttig ved flertydige ord for mere eller mindre umistelige legems/ejen-dele som arm, tunge og stol.

Ved kvantitative opgørelser over lemmaer behøver man næppe bekymre sig om den sidste rest af flertydigheder, men til en fuldt automatisk analyse kan man forudse, at der kræves både en ret gennemarbejdet syntaktisk analyse og et omhyggeligt udvalgt, større sæt af semantiske oplysninger.

## Noter.

(1) DANwORD, Hyppighedsundersøgelser i moderne dansk, ved Bente Maegaard og Hanne Ruus, se f.eks. SAML III, 4, 5.
(2) se. f.eks. Yorick Wilks: The Stanford Machine Translation Project, i Natural Language Processing ed. by Randall Rustin, New York 1973.
(3) se f.eks. Roger Schank: Identification of Conceptualizations Underlying Natural Language, i Computer Models of Thought and Language ed. by Roger C. Schank and Kenneth Mark Colby, San Francisco 1973.