23. februar 1990

RST

DET REDUCEREDE SEMANTISKE TRÆKSYSTEM

(Version 1.0)

af
Frede Boje, Carsten Hansen, Lene Schøsler og Ole Togeby

1. Formål

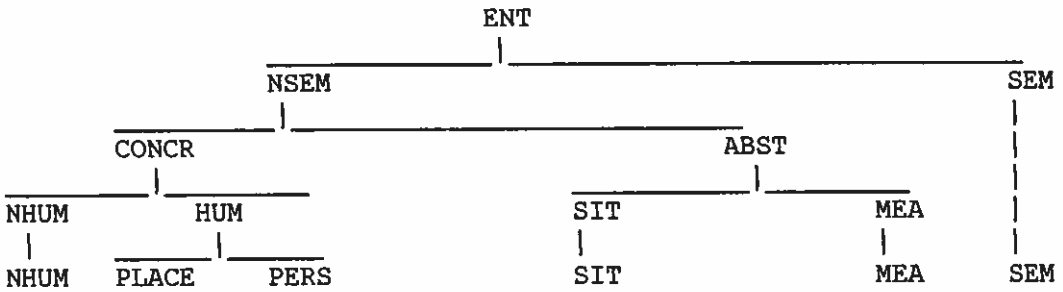
I det følgende beskrives det 'Reducerede Semantiske Træksystem' (RST) som har til formål at muliggøre disambiguering på semantisk grundlag: a) Alle substantiver kodes med værdier for semantiske træk, og b) alle prædikater (dvs. verber, adjektiver, præpositioner og rammebærende substantiver) kodes med angivelse af hvilke RST-værdier de selekterer i deres argumenter.

Ved hjælp af et sæt match-regler foretages der på IS følgende disambigueringer: 1) udelukkelse af de semantisk uacceptable blandt de (på ERS i analysen) syntaktisk genererede strukturer, 2) udelukkelse af de semantisk uacceptable kombinationer af readings af både det rammebærende prædikat og det kodede substantiv i hver prædikation i de genererede træer.

Featuresystemet definerer også både den øvre og den nedre grænse for antallet af readings et leksikalsk ord kan have. Kan to "reading-kandidater" ikke differentieres ved kodning med to forskellige værdier, må skellet mellem de to "readings" opgives. Finder man det rigtigst at kode et substantiv eller en argumentselektion ved to disparate værdier, er det tegn på at der er to readings af ordet.

2. Systemet

Featuresystemet ser således ud:



hus	USA	direktør	investering	måned	paragraf
bjerg	Europa	erhvervsliv1	handlingsprogram	tid	dokument
floder	Japan	virksomhed1	anvendelse	arten	forslag
	EF	Kommissionen	højteknologien	del	handlings-
		datamater	teknologi	slags	program
		industrien	elektronik	genop-	
		projektet1	initiativer	bygningsperiode	
			teknologiforskning		
			problem		
			iværksættelse		
			delmål		
			kvalitet		
			opdeling		
			udviklingen		
			arbejde1		
			metode		

Systemet består af 11 værdier, 6 terminale og 5 nonterminale.

Hvert substantiv skal kodes med én og kun én terminal værdi. Som nævnt i 1., er der tale om to readings af den leksikalske enhed, såfremt man ønsker at give to eller flere værdier til én leksikalsk enhed. Dette gælder dog ikke kategorien SEM; et ord der hører til denne kategori, hører samtidig både til en kategori under CONCR og en under ABST).

Hvert prædikat kodes for hvert af dets argumenter med en af de 11 værdier som det selekterer på den pågældende argumentplads.

3. Definition af de enkelte værdier

Konceptuelt er alle betegnelser (ENT, SEM...MEA) værdier at ét og samme træk. Implementeringsteknisk er det imidlertid nødvendigt at udforme systemet således, at navnene for alle nonterminale værdier samtidig er træknavne, således at træknavnet for en given værdi er moderknudens (værdi)navn. Hvis fx et ord (i en given reading) skal have værdien PERS, skal træknavnet være HUM, altså: {...hum=pers,...}. Den praktiske anvendelse af systemet beskrives nedenfor (afsnit 4).

a. SEM - NSEM: Et substantiv hører til i kategorien SEM (for semiotisk) hvis det ordet henviser til er noget der både kan 'læses', 'fortolkes', 'oversættes', og 'skrives' eller 'tegnes' - og altså optræde som A2 til disse verber.
Ord der ikke kan optræde som A2 for disse verber hører til kategorien NSEM (for nonsemiotisk).

Alle ord der hører til kategorien SEM vil også kunne optræde både som en underkategori til CONCR og som en underkategori til ABST. Ordet 'bog' er prototypisk for kategorien SEM. Den kan både optræde som NHUM, fx i 'hun kastede bogen efter ham', og som SIT fx: 'bogen om fødsler beskriver det hele'. Kategorien SEM er indført for at undgå opdelingen af alle semiotiske ord i to readings som alligevel i reglen ikke kan beregnes i mange kontekster.

b. CONCR - ABST: Et substantiv er CONCR (for konkret) hvis det betegner en fysisk ting eller stof der kan 'ses', 'høres' eller 'mærkes' og altså kan være A2 for disse verber.
Substantiver som ikke kan indgå som akkusativen i akkusativ med infinitiv-konstruktioner efter sanseverber, hører til kategorien ABST (for abstrakt).

Læg mærke til at sanseverberne kun udgør en test for CONCR når de følges af en akkusativ med infinitiv-konstruktion; eksemplet : 'hun så at toget var kørt' betyder nemlig ikke at hun så toget, men at hun 'forstod' at toget var kørt på grundlag af et eller andet andet som hun så, fx køreplanen, eller den tomme perron; man kan da heller ikke sige : *'hun så toget være kørt'. Ordet 'tog' hører til overkategorien CONCR fordi det kan indgå som akkusativen i eksemplet: 'Hun så toget køre'. Ord der hører til kategorien ABST kan også beskrives positivt som ord der har argumentstruktur (undtaget nomina agentis, som hører til PERS), eller som kan være grammatisk objekt for supportverber som 'foretage', 'gennemføre', 'gøre', eller som betegner tid, dimensioner eller mål.

c. HUM - NHUM: Et substantiv af kategorien CONCR er HUM (for human) hvis det betegner noget der kan handle intentionelt, dvs. være A1 for prædikater som 'gøre noget med vilje', 'holde op med at gøre det og det', eller være A3 for prædikatet 'tvinge nogen til at gøre det og det'.
CONCR-ord der ikke kan indgå som argumenter på disse pladser er NHUM (for nonhuman).

I kategorien HUM indgår ikke kun mennesker, men også maskiner og organisationer der kan handle, samt mange stednavne. På grund af den i teknisk politiske tekster meget udbredte metonymi, hvor stednavnet betegner 'folk på stedet', indgår visse geografiske stednavne i overkategorien HUM, fx 'Japan' i eksemplet: 'Japan har taget nye initiativer inden for teknologi'. Sådanne stednavne får den terminale værdi PLACE, medens stednavne, der ikke har mulighed for denne metonymi får terminalværdien NHUM, fx floder og bjerge.

I kategorien NHUM indgår ord der betegner de ikke humane levende væsener (fx 'husdyr', 'tyr', 'bakteriefloraen'), dele af mennesker og dyr (fx 'arm',

'blod'), andre naturlige ting og stoffer ('vand', 'flod'), alle forarbejdede ting ('hus', 'kredsløb', 'chip').

> d. PLACE - PERS: Et substantiv af kategori HUM er PLACE hvis det betegner et geografisk sted hvor man kan 'rejse hen til', 'tage afsted fra' og hvis det altså kan være A2 i forhold til disse prædikater. HUM-substantiver som ikke er PLACE hører til kategorien PERS (for person).

Til kategorien PERS hører ord der betegner kommunikationsmidler (fx: 'computer', 'datamat','radio'), organisationer (fx 'regeringen', 'erhvervslivet', 'befolkningen') og personer, herunder nomina agentis som har argumentstruktur (fx 'kongen', 'formanden', 'bageren').

> e. SIT - MEA: Et substantiv af overkategorien ABST hører til kategorien SIT (for situationel) enten 1) hvis det betegner en proposition (en mulig kendsgerning) og kan ekspliciteres ved en appositionel atsætning: 'den mulighed at han kom', eller 2) hvis det betegner en tilstand, et forløb eller en begivenhed i tid og det a) både kan være subjekt for mindst et af verberne 'ske', 'vare','foregå', 'finde sted' og b) har argumentstruktur, dvs. tager et argument med en bunden præposition, fx 'strukturen i materialet' = 'materialets struktur'. ABST-substantiver der ikke hører til kategorien SIT hører til kategorien MEA (for measure).

De fleste substantiver under kategorien SIT er afledte substantiver, dvs. substantiver afledt af verber eller adjektiver, fx 'anvendelse', 'behandling', 'kærlighed', 'vrede', 'investering'. Andre er ikke (danske) afledninger, men har, som afledninger, argumentstruktur og kan optræde som betegnelser for relationer, tilstande, processer eller begivenheder: 'uafhængigheden af ..', 'verdensproduktionen af kobber', 'kobberproduktionen', 'initiativ til', 'struktur i'. Nogle substantiver synes - ofte fordi de er komposita - kun at have en svag argumentstruktur, og de forekommer næppe som noget der har realitet i tid, men de kan slet ikke opfattes som konkret sanselige, de betegner discipliner eller brancher og regnes som SIT-substantiver: 'elektronik', 'teknologi', 'fællesskabsintervention'.
Kategorien MEA kan positivt beskrives som 1) betegnelser for tid (tidspunkter eller perioder), fx 'efterkrigstiden', 'periode', 'tid', eller 2) betegnelser for målelige dimensioner, som regel afledt af adjektiver, der kan indgå i konstruktioner som 'det og det af den og den XXX' og 'den havde en XXX på så og så mange YYY [måleenheder, se under 3)]', fx 'farve', 'længde','størrelse', 'omfang' (NB disse ord har altså argumentstruktur med to argumenter, men kan ikke 'foregå' eller 'ske' i tid), eller 3) betegnelser for måleenheder: 'grad', 'km', 'liter'. Sådanne måleenheder kan stå på YYYs plads i eksemplet ovenfor. Alle ord der på ERS har scat lig med SPECIFIER eller QUANTIFIER hører til kategorien MEA.

4. Kodningsvejledning

Kodningen af ord rummer både et indholdsmæssigt og et praktisk/organisatorisk aspekt, som ikke kan beskrives helt uafhængigt af hinanden. Alligevel er beskrivelsen her delt op i tre afsnit: 4.1: Hvordan opdeles i readings?, 4.2: Hvordan skal de færdige ordbogsindgange se ud?, 4.3: Hvordan kan kodnings-

arbejdet organiseres? Der er her regnet med at EDB endnu ikke kan bruges i kodningsarbejdet. Når/hvis det bliver tilfældet, vil vejledningen blive revideret.

4.1 Opdeling i readings

Udgangspunktet for arbejdet er indholdet af alle de ordbogsindgange i **/vaxusers/Lemma/ok**, der som lu-værdi har de ord som skal kodes. For hvert ord gøres følgende:

a) Undersøg, om ordet allerede er kodet med alle relevante betydninger, fx ved at slå ordet op i Vinterberg og Bodelsen: Dansk-engelsk ordbog og tage så mange af de der anførte betydninger som ikke kan udelukkes i vores teksttype. Det er fx et spørgsmål om betydningen 'målstolpe' tages med af ordet 'mål'.

b) Tag alle faste udtryk fra, og anfør dem i en særlig liste. Verber og præpositioner der kun kan tage et argument i nøgen form eller med arkaiske bøjningsformer, indgår som regel i et fast udtryk, fx 'de tabte mål og mæle', 'til lands'.

c) Medtag kun et skel mellem to betydninger af et ord hvis man kan lave en overbevisende zeugma. Dvs. Hvis man sideordner konteksten til de to betydninger som man undersøger, skal resultatet give en absurd eller komisk sætning. Eks.1: Ordet 'i' har både betydningen 'på tidspunktet X' og betydningen 'med en varighed af', for hvis man sideordner konteksten til disse betydninger får man en absurd eller komisk sætning: 'I begyndelsen og to dage skabte Gud ikke noget'. Eks. 2: prototypen på en zeugma er: 'Hvad er højest, Rundetårn eller et tordenskrald?'. Sætningen viser at ordet 'høj' både kan betyde 'som rager langt op' og 'som har kraftig lyd'. Eks. 3: Kan man skelne mellem betydningen 'med dimensionen X' og 'fordelt på' med konteksterne: 'en kjole i glade farver' og 'et hus i fire etager'? Nej. For det giver ikke zeugma at sideordne konteksterne: 'et hus i mange farver og fire etager'.

d) Tilføj readingsdistinktioner hvis en reading bedst kodes med to disparate træk. Eksempel: Hvis ordet 'mål' med fodbold-betydning medtages bør det kodes med både CONCR, fx 'de så målet knække på grund af stormen' og MEA fx: 'Svenskerne vandt med 7 mål mod 4'. Der er altså to fodbold-readings af ordet 'mål'.

e) Hvis ordet i to betydninger får samme kodning på alle kodningspladserne må de behandles som én reading. Eks: De to betydninger af 'i' der optræder i 'I begyndelsen skabte Gud ikke noget' og 'I fire dage skabte Gud ikke noget' må betragtes som samme reading (selv om det giver zeugma at sideordne dem) for RST-systemet kan ikke adskille deres kodninger.

4.2. Ordbogsindgangenes udseende

De færdige ordbogsindgange skal være **kodet i Lemma-format**, men med nogle tilføjelser i forhold til de hidtidige makroer, både i den "indre" (= den kørbare del) og den "ydre" ordbog.

4.2.1 Indre ordbog

4.2.1.1 Reading-nummerering

Hvis resultatet af reading-opdelingen (afsnit 4.1) bliver, at Lemmaordbogens readings (kørbare og/eller pseudo-) bevares, blot forsynet med de supplerende semantiske oplysninger, bibeholdes også reading-nummereringen. Hvis resultatet er en ændret opdeling, bevares de gamle reading-numre så vidt muligt, og nye readings nummereres fortløbende.

4.2.1.2 Nye features
Mellem **term=xx0** og **}** indføjes følgende:

a) For substantiver: RST-værdi med moderknudens navn som træknavn, altså fx **{..., hum=pers}** eller **{..., abst=mea}** eller **{..., ent=sem}**. Husk, at værdien altid skal være en terminal værdi. Hvis substantivet er prædikativt (= rammebærende), skal det desuden kodes i overensstemmelse med pkt. b):

b) For prædikater (verber, adjektiver, præpositioner og prædikative substantiver): De selektionskrav, der kan opstilles til argumenterne m.h.t. RST-værdier (terminale eller nonterminale). Hvis prædikatet fx har tre argumenter og arg1 skal være HUM, arg2 SEM og arg3 PLACE, skrives: **{..., a1concr=hum, a2ent=sem, a3hum=place}**. (Hvis oplysningerne om argumenternes eventuelle 'daparg' og 'pform' ikke allerede står i den foregående del af den indre ordbog, skal de selvfølgelig indsættes på den rette plads).

Specielt for adjektiver gælder, at deres selektionskrav til det substantiv, de modificerer (som attribut eller prædikativ), udtrykkes ved et træknavn med **'x'**. Hvis adjektivet fx kræver at subjektet er HUM, skrives: **{...xconcr=hum,...}**

4.2.2 Ydre ordbog

4.2.2.1 %% Coder + dato
For indgange hvor ændringen i den indre ordbog **kun** består i tilføjelse af RST-features, ændres denne linie ikke. Til gengæld tilføjes i så fald en ekstra kommentarlinie efter %% Examples: **%% Last update: [dato] by [navn].**

For indgange, som er nye eller har fået tilføjet eller ændret andre features (eller værdier), fx oplysninger om 'daparg', indsættes her det aktuelle navn + dato - fx ved hjælp af Lemma-makro (se nedenfor afsnit 4.3).

4.2.2.2 %% Source
I "uændrede" indgange bevares oplysningen. Ændrede og nye indgange forsynes med aktuel oplysning om kilde: Det opslagsværk, det følgende eksempel er taget fra, eller hvis eksemplet er hjemmelavet, koderens navn (eller initialer).

4.2.2.3 %% DEF
Her anføres 'glosse' på dansk og engelsk. Ved 'glosse' forstås her en definition af af readingens betydning(er) formuleret på en sådan måde at glossen kan indgå på samme syntaktiske plads i en sætning som ordet der defineres. Fx kan reading 2 af ordet 'mål' have glossen 'genstand for ens stræben'. Man kan altså i sætningen 'han nåede sit mål' med lidt god vilje indsætte glossen på ordet 'mål's plads: 'han nåede genstanden for sin stræben'.
Den engelske oversættelse skal være et og kun et ord (eller en og kun en fast forbindelse af ord) således at det engelske ord er den korrekte oversættelse i alle eksempler hvor det danske ord forekommer i den pågældende reading. Kan man ikke finde et engelsk ord der kan det, er det tegn på at der er tale

om to readings på dansk. Kun hvis man på dansk med RST ikke kan definere forskellen på de to betydninger, kan man anføre to engelske ord som oversættelse af en dansk reading.

4.2.2.4 %% Comments
Her er der forholdsvis frit slag. Husk, at af hensyn til kompileringen må kommmentarer ikke fylde mere end én linie. Hvis der er behov for længere kommentarer, må der sættes nye kommentartegn (%%) for hver linie.

4.2.2.5 %% Examples
For alle indgange skrives mindst ét eksempel. (Husk %% ved ny linie). Eksempler skal være kontrasterende og der skal være eksempler på alle ARGUMENTPRÆPOSITIONER. At eksemplet er kontrasterende vil sige at ordet forekommer i en kontekst hvor det kun kan have en af de betydninger som den leksikalske størrelse i alt kan have. Hvis man fx skelner mellem: 'nå_v1: komme (hen) til' og 'nå_v2: gøre noget der fører til målet X', kan man ikke anføre eksemplet 'han nåede det', for her kan man ikke se hvilken af betydningerne der er på spil.

4.3 Organisering

Som nævnt ovenfor, skal alle færdige ordbogsindgange være kørbare Lemma-ordbogsindgange. Det betyder ikke, at man **skal** bruge Lemma-makroerne til kodningen, men det vil nok normalt være det mest praktiske. Man kan **for eksempel** bruge følgende arbejdsgang:

1) Man opretter en midlertidig arbejdsfil i /vaxusers/Lemma/work.

2) Man henter kopier af alle de ordbogsindgange i /vaxusers/Lemma/ok, der som lu-værdi har de ord, som skal kodes. Det kan fx gøres ved at bruge makroen **ind**. (Husk, at den skal loades; se vejledning uddelt på ordbogsdagen (10. maj 89) eller spørg Frede).

3) Under arbejdet med den ønskede reading-opdeling (afsnit 4.1 ovenfor) gøres de relevante notater om værdier og eksempler i frit format i arbejdsfilen eller på papir.

4) Med disse notater på skærmen eller papiret kodes nye indgange med brug af Lemma-makroer, hvorefter de supplerende RST-oplysninger indføjes.

5) De færdige indgange kompileres og "afleveres" til den ordklasse-ansvarlige.