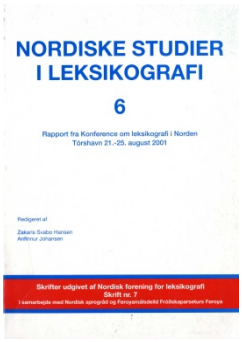
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Sprogteknologiske ordbaser for de nordiske sprog – rapport fra et forskningsnetværk

Forfatter: Bolette Pedersen, Ruth V. Fjeld & Maria Toporowska Gronostaj

Kilde: Nordiske Studier i Leksikografi 6, 2003, s. 273-290
Rapport fra Konference om leksikografi i Norden, Tórshavn 21.-25. august 2001

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Bolette Pedersen, Ruth V. Fjeld & Maria Toporowska Gronostaj

Sprogteknologiske ordbaser for de nordiske sprog – rapport fra et forskningsnetværk

This paper deals with a Nordic research network, SPINN, concerned with the harmonisation of Nordic, semantic computational lexica. Plans for extension of the lexica are described and some problems in adapting the Danish lexicon into a Norwegian version are accounted for. The main equivalence problems are caused by semantic changes which are taking place in modern Norwegian. These are changes with concequences for both internal and external lemma selection and especially the argument structure. Finally, a possibility to establish a multilingual network for Danish, Norwegian and Swedish is considered. The languages in the network will be interlinked by means of English, as described in The New Oxford Dictionary of English (NODE). Attention is also paid to some polysemy-oriented procedures supporting the interlingual linking.

1. SPINN – Sprogteknologi og INformationssøgning i Norden

1.1 Formål og baggrund

Intentionen med det tværfaglige netværk SPINN (SProgtekonologi og INformationssøgning i Norden, http://www.cst.dk/spinn/spinnhome.html) bevilget for en 2-årig periode af Nordisk Ministerråd, er at sikre at de forskningsmiljøer i Skandinavien og i Island der arbejder med opbygningen af ordbaser egnet til sprogteknologiske applikationer, samt de forskningsmiljøer der arbejder med indholdsbaserede søgemaskiner, udveksler erfaringer og indleder et fagligt samarbejde. Mere konkret er målet på længere sigt at gå sammen om at etablere den nødvendige basis for en sammenhægtning af de nordiske sprogteknologiske ordbaser således at også tværsproglige applikationer mellem de nordiske sprog kan komme på tale.

Det faglige udgangspunkt for iværksættelsen af SPINN-netværket udgøres af det dansk-svenske samarbejde der blev indledt med ordbaseprojektet SIMPLE (Semantic Information for Plurilingual, Multifunctional Lexica, se Lenci *et al.* 2000). I dette EU-projekt har man skabt forudsætningerne for at udarbejde semantiske, sprogteknologiske ordbaser for 12 sprog med hver især 10.000 ordbetydninger: 7000 substantiver, 2000 verber og 1000 adjektiver. For hver ordbetydning angives informationer af typen (i) begrebstype (semantisk klasse) hentet fra SIMPLE-ontologien som består af 139 hierarkisk ordnede top-ontologibegreber, (ii) domæne, (iii) betydningsdefinition, (iv) teksteksempel, (v) argumentstruktur (semantisk valens), (vi) selektionsrestriktioner, altså hvilke semantiske krav der stilles til ordets argumenter (f.eks. humant subjekt), (vii) semantiske relationer og semantiske træk, samt endelig synonymi-, polysemi- og kollokationsrelationer.

1.2 Netværkets aktiviteter

Netværket afholder møder og seminarer med det formål at komme nærmere et egentligt samarbejde med henblik på harmonisering af ordbaser og deres anvendelse til bl.a. søgning

(Pedersen, Toporowska & Fjeld, 2001 og 2002). For at muliggøre dette præsenterer netværkets deltagere først og fremmest deres ordbaser og ordbaseværktøjer for hinanden og diskuterer hvilke typer data der er relevante i sprogteknologiske ordbaser eksempelvis i forhold til sprogindustrien og med særligt fokus på informationssøgning.

Deltagerne i netværket inddrager ydermere det tværsproglige aspekt og foretager konkrete sammenligninger mellem ordbogsudtræk fra forskellige nordiske sprogteknologiske ordbøger, bl.a. SIMPLE og NordKompLeks. Formålet er at se på forskelle og ligheder i strukturering og detaljeringsgrad af leksikalske informationer, såsom ontologisk tilskrivning, valens, argumentstruktur og selektionsrestriktioner.

Endelig inviterer netværket eksperter fra udlandet til at orientere om status inden for udviklingen af flersproglige, sprogteknologiske ordbaser samt om deres erfaringer med evaluering af forskellige tværsproglige søgesystemer. Sproglige tilgange til informationssøgning er nemlig i skarp konkurrence med mere statistisk baserede metoder og det er derfor vigtigt at belyse hvordan sprogteknologien kan udnyttes yderligere. F.eks. bør man overveje hvorvidt sprogbaseret informationssøgning kan fokusere mere på den måde hvorpå søgeresultaterne præsenteres og dermed udnytte den sproglige tilgang ikke alene i analysen og opmærkningen af tekster, men også i høj grad i formidling af søgeresultater.

1.3 Videreudvikling og sammenhægtning af ordbaser

Netværkets aktiviteter skaber en god baggrund og et godt forum for videreudvikling og sammenhægtning af de sprogteknologiske ordbaser. Specielt de praktiske øvelser med at sammenligne ordbeskrivelser i de forskellige eksisterende ordbaser giver en fornemmelse af hvilke problemstillinger et sådant harmoniseringsprojekt vil blive konfronteret med. Udover at ordbaserne ikke overraskende indeholder forskellige *typer* af leksikalsk information fordi de i nogen udstrækning er opbygget til forskellige formål, så kommer særligt to aspekter i fokus under denne 'øvelse':

- Ordbaserne udviser store forskelle mht. dækningsgrad hvilket vanskeliggør sammenhægtning; det er f.eks. højst forskelligt hvor mange valensmønstre der er beskrevet for de enkelte verber; for nogle ordbasers vedkommende er kun de 'prototypiske' medtaget, hvor andre er mere komplette. Dette skyldes bl.a. at der for stort set alle disse basers vedkommende er tale om 'pilotprojekter', dvs. mindre ordbaser hvor formålet primært har været at afprøve forskellige beskrivelsesmodeller. Forskellige aspekter af dækningsgrad har derimod været mindre centralt for disse projekter.
- Specificeringsgraden er forskellig. Dette problem kommer særligt tydeligt frem i forbindelse med de semantiske beskrivelser i de danske og svenske SIMPLE-ordbgøer, hvor to forskellige anskuelsesvinkler kan tages: man kan vælge at give meget specifikke semantiske oplysninger for f.eks. verbalbetydninger og deres selektionsrestriktioner. Dette giver mange læsninger med deraf mulige problemer med entydiggørelse. Eller man kan vælge at anskue verbers underbetydninger mere generelt med beskrivelser der rummer flere generelle antagelser. Dette giver muligvis fordele mht. entydiggørelse; til gengæld giver det mindre præcise beskrivelser.

I det følgende vil vi komme nærmere ind på disse og relaterede problemstillinger. I afsnit 2 gives en beskrivelse af det kompletteringsarbejde der forestår med den danske SIMPLE-ordbog, hvor et større korpusarbejde er i opstartsfasen. Afsnit 3 beskriver arbejdet med den

norske SIMPLE-ordbog hvor udviklingen i sproget afstedkommer en diskussion af hvad der bør dækkes af den sprogteknologiske ordbase. Afsnit 4 beskriver vore overvejelser vedrørende en sammenhægtning af ordbaserne med udgangspunkt i *The New Oxford Dictionary of English* (NODE).

2. Komplettering af den danske SIMPLE-ordbog

2.1 Dækningsgraden i SIMPLE-DK

Den danske SIMPLE-ordbase (http://www.cst.dk/simple/index.html) var et forsøgsprojekt hvor intentionen primært var at få afprøvet den såkaldte SIMPLE-model (Lenci et al. 2000) på det danske sprog i et omfang af 10.000 betydningsbeskrivelser. Fokus blev lagt på beskrivelsen af substantiver, idet disse som nævnt skulle udgøre mere end to tredjedele af ordbogen. I den danske gruppe bestræbte vi os på at dække SIMPLE-ontologien jævnt, dvs. at få repræsenteret alle ontologiske typer helst med mindst 100 eksempler[1]. Det betyder at det ikke udelukkende var hyppighed der blev den bestemmende faktor for hvilke ord der blev udvalgt. Da SIMPLE-projektet var en videreudvikling af det morfologiske og syntaktiske ordbaseprojekt PAROLE (Ruimy et al. 1998), var udgangspunktet for ordbasens dækningsgrad ganske vist givet på forhånd. PAROLE-ordforrådet dækkede 20.000 ord og var baseret på frekvensundersøgelser af Den Danske Ordbogs Korpus på 40 mill. løbende ord. Frekvensundersøgelserne blev foretaget på et tidspunkt hvor automatisk ordklasseopmærkning ('POS tagging') endnu ikke var færdigudviklet for dansk. Derfor var frekvenslisten noget fejlbehæftet idet f.eks. sammenfaldende ordformer i form af absolut homografi kunne give fejlagtige 'udsving' på frekvenslisten. F.eks. har mange verber i nutid sammenfald med deres afledte nomen agentis; så som *løber* og *aftager*. Det faktum at korpus ikke var ordklasseopmærket gjorde således at mange ikke særligt hyppige verbalsubstantiver fik en fejlagtig høj frekvens fordi de ikke kunne adskilles fra nutidsverbalformen og således kom med blandt de 20.000 hyppigste ord i dansk. Sådanne 'udsving' i PAROLE-ordbogen sammenholdt med princippet om at få afprøvet SIMPLE-modellen bedst og bredest muligt gør at SIMPLE-DK *ikke* dækker de 10.000 hyppigste betydninger i dansk men kun et udsnit af disse. Udover behovet for generelt at få udvidet ordforrådet i SIMPLE-DK er der altså også tale om nogle helt klare 'huller'; dvs. man vil givet kunne nævne en del meget hyppige ord som ikke er at finde i den nuværende SIMPLE-ordbog og som derfor mangler for at ordbogen kan finde praktisk anvendelse.

Dækningsgrad kan også anskues ud fra en anden, mere semantisk vinkel; dvs. vi kan tale om hvorvidt et ord er dækket rent semantisk i højere eller mindre grad. Her vil man i SIMPLE-DK se nogen forskel de enkelte ordklasser imellem. For substantivernes vedkommende er vi er meget tæt på en 'fuld' semantisk dækningsgrad idet stort set alle de kodede substantiver er kodede i alle deres betydninger (baseret på Nudansk Ordbog's betydningsbeskrivelser). For adjektivernes og verbernes vedkommede er dette langt fra tilfældet. For at få dækket alle de ontologiske typer der er relevante for disse ordklasser og for f.eks. at få afprøvet forskellige typer af selektionsrestriktioner ved verberne, er mange af disse ord ikke nødvendigvis kodet i alle deres betydninger. Det gælder især for de højfrekvente verber som kan optræde med virkelig mange betydninger, hvoraf mange ikke engang er beskrevet i en mellemstor ordbog som Nudansk Ordbog. Specielt de forskellige partikelkonstruktioner som ofte giver en helt ny betydning af verbet, er ikke kortlagt i nogen eksisterende ordbog for dansk.

Dette kan have flere årsager. For det første er de eksisterende ordbogsressourcer for

dansk ikke alle lige baserede på korpora, dvs. der kan være vigtige og temmelig frekvente betydninger som simpelthen er undgået ordbogsredaktørens opmærksomhed. En anden forklaring kan være at vi her har fat i en af forskellene mellem en 'traditionel' ordbog med mennesket som bruger i modsætning til den sprogteknologiske ordbase med computeren som primær 'bruger'. Den traditionelle ordbog har et formidlingsaspekt hvor det er vigtigt klart at få formidlet det prototypiske omkring ordets betydning. Ved hjælp af denne prototypiske definition kombineret med en række eksempler kan det i mange tilfælde overlades til brugeren selv at 'ekspandere' betydningen til også at dække nært beslægtede betydninger. Dette er langt vanskeligere i datamatisk sammenhæng hvor man oftest er tvunget til at være fuldstændig eksplicit (Nimb & Pedersen 2000). Hvor man i en ordbog som Nudansk Ordbog kan komme afsted med at tolke *han passerede de 70* som et eksempel der kan gå ind under definitionen *gå, rejse el. på anden måde bevæge sig forbi el. igennem nogen el. noget*, så ville man i en sprogteknologisk ordbase være tvunget til at definere dette som en omend relateret, så dog meget anderledes tidslig betydning med en anden selektionsrestriktion på objektet, nemlig et krav om at objektet skal have en tidslig dimension (f.eks. *70 år*) og at der i denne betydning af *passere* ikke foregår *gang* eller *rejse* i ordets egentligste forstand. Vi er med andre ord tvunget til at tage begrebet semantisk dækningsgrad meget bogstaveligt i den sprogteknologiske sammenhæng, hvilket selvfølgelig gør at kompletteringsarbejdet bliver temmelig omfangsrigt. Til gengæld kan vi i en sprogteknologisk ordbase vælge at sætte grænsen et andet sted; idet vi kan nøjes med at tage de *statistisk* signifikante betydningsnuancer med ved et ord. Dette er ikke altid hensigtsmæssigt i en traditionel ordbog.

2.2 Korpusbaseret komplettering af den semantiske dækningsgrad ('Senseval-metoden')

I forbindelse med et semantisk opmærksningsprojekt kaldet Senseval[2] (http://www.cst.dk/senseval/index.html), har vi på CST undersøgt 100 ord med henblik på hvor meget korpusmateriale det var rimeligt at inddrage for at få dækket statistisk signifikante betydninger af disse ord. Dette arbejde ligger til grund for den igangværende komplettering af den danske SIMPLE-ordbase.

Der anvendes følgende algoritme for korpusudtræk for et ord der skal beskrives i ordbasen:

- Ordet i alle dets relevante bøjningsformer slås op i korpus.
- Hvis ordet forekommer i flere end 115 eksempler, foretages en reduktion af konkordansen efter følgende princip (adapteret fra Senseval-projektet): (i) For hvert lemma tælles antallet af betydninger i fx Nudansk Ordbog (eller en ordbog af tilsvarende størrelse), underbetydninger og idiomatiske udtryk tælles med. (ii) Det antal korpuseksempler der skal undersøges, vil være 100 + 15*n* hvor *n* står for antallet af betydninger i ordbogen. Konkordansen sorteres på højresiden for at undersøge mulige valensled. Ved substantiver kan det muligvis være en fordel også at sortere til venstre for at opdage evt. genitiver.
- De syntaktiske mønstre der optræder i korpusmaterialet medtages i den semantiske beskrivelse, men en vis grad af introspektion er dog nødvendig for at dække fejl og statistiske huller.

Hvis vi f.eks. slår et verbum som *røre* op i Nudansk Ordbog, finder vi følgende beskrivelse:

1. røre (ved) ng(t) bringe hænder el. fingre i kontakt med noget = BERØRE du må ikke røre ved de udstillede ting, han rørte hende let på armen
røre ngt bevæge noget d røre benene
røre (på) sig bevæge sig d han kan ikke røre sig, rører sig ikke ud af stedet, vi trænger til at komme ud og røre os, ikke en vind rører sig, de sovende begyndte så småt at røre på sig
2. røre ngt (dagl.) spise el. drikke noget, jeg rører aldrig fisk, hun rører ikke spiritus
3. røre ved ngt (dagl.) beskæftige sig med noget, jeg vil ikke røre ved den sag, forfatteren rører her ved et stort problem
4. røre i ngt bevæge noget rundt i noget, fx en ske rundt i noget, røre i farsen, røre i malingen
røre ngt {op med/ud i} blande noget med noget andet farsen røres op med mælk, pulveret røres ud med vand
5. røre ng bevæge nogen følelsesmæssigt = GRIBE hendes sorg rørte mig dybt, jeg blev dybt rørt

Når underbetydninger og idiomatiske udtryk tælles med, giver det en faktor på 8. Man skal altså udtrække 100 + (15 x 8) = 220 korpuseksempler for med rimelighed at 'dække' dette ord, ifølge den opstillede metode.

Ved at følge dette princip får vi set fra en rent syntaktisk synsvinkel følgende resultater for et verbum som *røre*, altså 12 syntaktiske mønstre og 3 idiomatiske udtryk:

EKSEMPEL	**SYNTAKTISK MØNSTER**
nok se, ikke røre	SUBJ verbum
derfor rører jeg intet når jeg skal på scenen	SUBJ verbum OBJ
de skal have handsker på når de rører ved træet	SUBJ verbum POBJ (*ved*)
pestoen røres tynd med noget af vandet fra pastaen	SUBJ verbum OBJ ATROBJ (*tynd/tyk..*)
det værk der rører ham dybest	SUBJ verbum OBJ MANNERADV (dybt)
en stor træslev der rører rundt..	SUBJ verbum-*rundt*
hvad der rører sig i befolkningen	SUBJ verbum-*sig*
det er svært at røre ud	SUBJ verbum-*ud*
jævn med kartoffelmel rørt ud i koldt vand	SUBJ verbum-*ud* POBJ (*i*)
æggeblomme rørt op med en spsk. vand	SUBJ verbum-*op* OBJ POBJ (*med*)
creme-fraiche rørt med dijonsennep	SUBJ verbum OBJ POBJ (*med*)
en ostecreme rørt af gorgonzola	SUBJ verbum OBJ POBJ (*af*)
røre på sig	IDIOMATIC
uden at røre en finger	IDIOMATIC
rørte vande	IDIOMATIC

Set fra en semantisk synsvinkel får vi identificeret 3 nye betydninger og 2 nye idiomatiske udtryk som ikke var beskrevet i Nudansk Ordbog:

- ændre på noget i negativ retning: *vi vil ikke røre ved børnepengene*
- foregå: *det er vigtigt at vide hvad der rører sig i befolkningen*
- skade: *kommunen kan ikke røre ham*

- idiomatisk udtryk: *uden at røre en finger*
- idiomatisk udtryk: *rørte vande*

Da disse er ganske frekvente og med en betydning der afviger ganske betragteligt fra verbets grundbetydning, bør der ikke være tvivl om at disse bør repræsenteres i ordbasen. Specielt 'foregå'-betydningen var meget markant i korpus.

3. Problemer ved utvikling av et norsk SIMPLE-leksikon

3.1 Arbeidsplan

Med utgangspunkt i det danske SIMPLE-leksikonet har vi begynt å lage en norsk parallellversjon med norske ekvivalenter og belegg fra Tekstlaboratoriets Oslokorpus. De første og avgjørende spørsmål i tilpassingen er å finne ut hvilke lemmaer som velges for det norske leksikonet, og hvilke lemmaer fra den danske basen som skal få en ny definisjon i den norske.

3.2 Ytre lemmaseleksjon

Det danske lemmautvalget er dels gjort ut fra den felleseuropeiske basen, dels for å dekke de forskjellige ontologiske typene i SIMPLE. Det må vurderes om denne ontologien er passende for norsk språk og kultur, eller om det må legges til hele nye begrepsfelt og/eller nye enkelttermer. For eksempel er begrepsfelt som *fjell* og *skisport* viktig i det norske språket. Enkeltuttrykk som *bindingsverkshus* og *tinnsoldat* er mindre aktuelle ord i norsk enn *lutefisk* og *oljeplattform*. Lemmalistene i ordbøker og ordlister skal danne utgangspunktet for lemmautvalget. Det skal suppleres med frekvensstudier av et helt nytt bokmålskorpus med tekster fra 1985 til i dag, et eksemplarisk korpus av litterære tekster, sakprosa og faglitteratur (Runde 2000). Danske lemmaer som ikke kommer med her, vil fortsatt være med i basen med den danske definisjonen. Retting av dem blir foretatt dersom det ved bruk av leksikonet viser seg å være diskrepans mellom norsk og dansk. Leksikonet skal altså være åpent for endringer og tillegg etter hvert som behovene avklares.

3.3 Indre lemmaseleksjon

Hvilke informasjonstyper vi velger å legge inn for hvert lemma, blir bestemt av tilgjengelig bakgrunnsmateriale fra tidligere analyser samt av behovsanalyser for det nye leksikonet, men vi prøver prinsipielt å legge det norske leksikonet så nær opp til de danske og svenske basene som mulig.

Arbeidsgangen videre er å kontrollere de danske definisjonene i forhold til de norske ekvivalentene og undersøke om de er i overensstemmelse med definisjonene i Bokmålsordboka. I neste omgang må det kontrolleres om definisjonen i Bokmålsordboka er i samsvar med belegg fra det nye bokmålskorpuset. Når fridefinisjonene er korrigert, kan vi gå gjennom de danske formaliserte semantiske beskrivelsene og justere dem for norsk.

3.4 Eksempel: Betydningsbeskrivelse av verbet "knuse"

Justering av betydningsbeskrivelsene er et ressurskrevende arbeid, som her skal illustreres med eksempler fra definisjonen av noen verb. Verb er valgt siden vi der kan støtte oss på NorKompLeks (Nordgaard 1996), der beskrivelsen av verbenes agrumentstruktur er formalisert. Denne beskrivelsen er et godt utgangspunkt, men definisjonene er gjort bare med eksemplene i Bokmålsordboka som materiale, og kan derfor ikke regnes som fullstendig

for beskrivelse av moderne norsk. Et hovedproblem er at ordbokseksemplene ikke er valgt ut spesielt for å illustrere verbenes argumentstruktur, de skal fylle mange andre funksjoner også. Korpussøk der argumentstrukturen studeres spesielt, vil sannsynligvis gi en del andre resultater for flere av verbene.

Studier av verbsemantikk i moderne norsk tyder på at det har skjedd en del endringer i argumentstrukturen for flere sentrale verb (jf. Sveen 1996). Det ser ikke ut til at den samme utviklingen er skjedd i dansk og svensk, noe som kan tyde på at norsk er det nordiske språket som har blitt mest påvirket av engelsk språkstruktur. Det har blant annet medført en del endringer i verbs transitivitet.

Ved Seksjon for leksikografi og målføregransking ved Universitetet i Oslo har man gjennom vel 30 år samlet belegg på endringer i norsk bokmål. Beleggene er søkbare i en base kalt **Nyordsmaterialet** (http://www.dokpro.uio.no/bokmaal/nyord/nyord_fside.html). Funn i denne basen bekrefter Sveens konklusjoner.

Verbet *knuse* er et av de verbene som har endret argumentstruktur. I den danske SIMPLE-basen har det følgende beskrivelser:

```
<SemU
  id="USEM_V_knuse_CCS_1"
  naming="knuse"
  example="Drengene ville skaffe sig adgang til ejendommen ved at knuse et
vindue med sten"
freedefinition="få noget til at gå i stykker el. i opløsning (NDO)"
WVSFTemplateCauseChangeofStatePROT
WVSFTemplateSuperTypeCauseRelationalChangePROT
WVSFEventTypeTransitionPROT
TSVP_CHANGE_TS_classificateur_de_verbe">

<Argument id="ARG1PREDknuse_CCS_1"
  comment="the first argument of the predicate knuse"
  semanticrolel="Role_Underspecified"
  informargl="ArgHumanAnimal">

<Argument id="ARG2PREDknuse_CCS_1"
  comment="the second argument of the predicate knuse"
  semanticrolel="Role_ProtoPatient">
```

På menneskespråk vil det kunne tolkes slik at en levende aktør (ArgHuman Animal) må utføre den handlingen verbet betegner slik at den går ut over et objekt (Role_ProtoPatient) som innebærer en endring i et objekts tilstand (Cause ChangeofState).

I det foreløpige norske leksikonet har vi hittil bare lagt til et norsk belegg. Formen er identisk, som i et stort antall av de danske lemmaene:

```
naming="knuse"
exampleN="Uvedkommende banet seg adgang ved å knuse et dørvindu i bakgården"
```

Definisjonen i Bokmålsordboka er "slå, klemme, male i stykker", og den skal legges inn i en norsk fridefinisjon, men i tillegg bør det i den tydeliggjøres at verbet har et kausativt betydningselement, slik som på dansk, "få noe til å gå i stykker". Slik kan utbygging av den norske SIMPLE-varianten også bidra til å forbedre Bokmålsordboka.

Beskrivelsen i NorKompLeks-basen ser slik ut:

w(knuse,30604,[trans1]).

Her er altså verbet koda som et vanlig, transitivt verb basert på beskrivelsen i Bokmålsordboka.

Men Nyordsmaterialet gir følgende belegg:

> - En 60 år gammel mann fra Oslo slår et ølglass i bakhodet på en annen mannlig gjest på restaurant "Schrøder" i Oslo. Slaget er så kraftig at **ølglasset knuser**.
> - En typisk nordisk skikk var å klinke egg. Man hardkokte for eksempel to egg, og to personer fikk hvert sitt. Så slo man eggene mot hverandre, og hvis et av **eggene knuste**, hadde vedkommende tapt.

I de nye eksemplene er kausativelementet borte og verbet er dermed ikke lenger transitivt. Hvis det språkteknologiske leksikonet skal brukes i søk mot unormert språk, noe vi ser som ønskelig, er det helt nødvendig å oppdatere den leksikografiske beskrivelsen slik at mange tidligere transitive verb også kan tolkes som et intransitivt verb, da dette er konstruksjoner som er svært vanlige, både i muntlig språk og i romaner og annet mer privat skriftspråk. For eksempel hører man på T-banene i Oslo en høyttalerstemme som sier hver gang et tog skal til å gå*: "Se opp for dørene, dørene **lukker**!"*

Vi må altså i den norske SIMPLE-basen opprette en artikkel **knuse II**, der verbet kan stå med bare ett argument.

3.5 Generelle endringer i norske verbs argumentstruktur

De endringene i transitiviteten som er registrert i Nyordsmaterialet, kan klassifiseres i tre typer: verb som kan ha et argument mer enn før, verb som mister et argument, og verb som får knyttet til seg en nærmest fast partikkel og derved også kan ha et argument til. I beleggsamlingen nedenfor er årstall for belegget tatt med, noe som viser at endringen ikke bare er et moteblaff i norsk språkutvikling:

3.5.1 Verb som kan ta et ekstra argument:

Klart kausative verb:

1. Damen i butikken blir ganske sikkert skremt for han oppfører seg som en villmann og **drypper** blod overalt. 1973.
2. Vi **dumper** ikke prisene, sier direktør <NN> i Norol. 1979.
3. Gemini Consulting, som i fjor **este** sin Norges-omsetning fra 70 til 210 millioner kroner. 1996
4. Hva **fenger** yngre sivilingeniører? 1975.
5. Briterne [- - -] **eksploderer** H-bombe til 2 milliarder kr. 1957.
6. Israel **festner** taket på de erobrede områdene. 1972.

7. Her holder Utah Bulwarks tidligere trener og kusk Stig H. Johansson et godt tak på hesten han **glapp** etter oppgjør med eieren Tony Malmgren. 1987.
8. Reglementet forbyr imidlertid noen å "**gro**" skjegg i løpet av tjenestetida, og sier dessuten at håret skal være kortklipt. 1970.
9. Tenk deg å få dovremakko i geitosten og smør i trusa , **grøsser** Anne Marit Jacobsen. 1999.
10. Det ville **klarne** diskusjonen hvis aksjeselskapene med tilslutning fra publikum ble skilt ut som en egen gruppe. 1968.
11. Dessverre har vi **kollidert** stevnet med NM i kl. B. 1969.

Uklart om verbet er kausativt:

12. Vi får flere og flere utlendinger som **ferierer** påsken i Nord-Norge. 1969.
13. Etter et år som sjef i KS, **flagger** Thorsdal at bedre bruk av informasjonsteknologi skal bli organisasjonens flaggsak. 1993.
14. Lærerne **flykter** skolen. 1988.
15. Bjørn Dæhlie og kompani **flykter** Norge for å få mer ro. 1996.
16. Pappenheimer har i åtte år ledet en gruppe som har **forsket** søvn ved Harvard. 1975.
17. Samtlige arbeidere i Berg Seterskog har selvfølgelig full anledning til å benytte seg av vår tillatelse til å **jakte** småvilt. 1970.
18. Derfor er det vanskelig å få en same til å **joike** en avdød, her møter man en slags tabu. 1956.

Etablering av nye verb:

Der det blir såkalte nye verb, får vi en slags parverb, dvs. to verb med samme semantikk men med forskjellige syntaktiske egenskaper. Tradisjonelt har vi jo hatt mange av disse i norsk, men betydningen har vært markert med forskjellig bøying:

henge - hengte - hengt = transitivt
henge - hang - hengt = intransitivt

Mens det intransitive *kjøre* og det transitive *kjøre* har nøyaktig samme bøying.
Skillet er i ferd med å forsvinne i moderne norsk:

19. Politimannen slang avisen rett foran meg (Aftenposten 26.4.2001).

Materialet gir følgende belegg:

20. Vi skal **kjøre** en redelig og ordentlig politikk. 1992.
21. Han ville ha tall på bordet, som bekreftet annonsens evner til å **kommunisere** budskapet. 1979.

3.5.2 Verb som får tillegg av partikkel uten endring i argumentstruktur:

22. Man **endrer på** reglene. 1972.
23. Om vi xx **fokuserer på** det mangfold av oppgaver som utføres innen husholdet. 1980.
24. Det er svært vanskelig eller nær sagt umulig å **forebygge mot** angrep fra lakselus. 1990.
25. Stoler vi ikke på Folketrygden går vi til privat forsikring eller **forhandler frem** bedre tjenestepensjoner i bedriften. 1991.

26. Har det aldri falt kiropraktorene inn at deres spesielle metode, om den skulle bli akseptert, bare utgjør en liten del av det behandlingsmønster som seriøst **forskes fram** i medisinen. 1974.
27. Før jeg går videre xx vil jeg kort **kommentere på** et problem som angår denne argumentasjonsmåtens xx begrensning. 1985.

Når et verb slik får knyttet en partikkel så fast til seg, kan det skje at vi får utviklet et nytt verb som ikke er transitivt i seg selv, slik det har vært tidligere. Det er en allmenn tendens i norsk, men den er lite beskrevet hittil.

3.5.3 Verb som mister et argument:
Nedenfor vises noen verb som er blitt **ergative**, det vil si at de er intransitive med et subjekt som er det samme som objektet for det tradisjonelt transitive:
28. Kinopublikum **flokker** til kinoene 1983.
29. Vil det lønne seg å **gjenvinne?** 1976.
30. Isflakene **knuger** mot hverandre. 1987.
31. Reparér før ruten **knuser.** 1993.
32. Slaget er så kraftig at ølglasset **knuser**. 1995.
33. En del av flaskene **knuste** når de nådde vannet. 1975.
34. Dagens doktorand kan **knytte** til den teorien. 1960.

3.5.6 Verbet får tillegg av adjektiv eller annen utfylling:
35. Helsetrøya **fisker** veldig godt i elver. 1993.
36. Han **fotograferer** så godt 1975.

Endringene er sannsynligvis konstruksjonslån fra engelsk, og er etter hvert blitt så vanlige at de fleste språkbrukere under 30 år ikke klarer å oppdage dem. Det er et tegn på at de går inn i norsk språkstruktur og bør være med i beskrivelsen i et norsk språkteknologisk leksikon. Det samsvarer for øvrig med de krav Lenci et al (2000) skisserer for verbbeskrivelse i SIMPLE-leksikonet. Om den bør være med i et normativt og pedagogisk leksikon, er en diskusjon som ikke kan føres her.

4. Mot en flerspråkig databas

I de föregående avsnitten har vi berört problematiken kring formaliseringen av ordbetydelser i lexikaliska databaser för danska och norska, vilket antyder vilken typ av frågor som kommer att aktualiseras inom en flerspråkig lexikalisk databas för skandinaviska språken. Nedan skisserar vi en modell för upprättandet av en sådan flerspråkig lexikalisk databas med utgångspunkt och stöd i enspråkiga databaser samt maskinläsbara tvåspråkiga lexikon. Vi föreslår en interlingua-baserad modell för sammanlänkningen av ordbetydelser i de skandinaviska språken (avsnitt 4.2) samt ger några praktiska genvägar till etableringen av ekvivalentpar bestående av en källspråksenhet och dess ekvivalent i engelska, vårt interlingua-språk (avsnitt 4.3). (Beskrivningen av modellen i 4.2 har även ingått som en delavsnitt i Pedersen, Fjeld, och Toporowska Gronostaj 2001).

4.1 Bakgrund

Det inom SIMPLE-projektet initierade arbetet med framtagning av språkteknologiska lexikon för 12 språk, inklusive lexikonmoduler för svenska och danska, samt det pågående arbetet med att bygga upp en norsk språkteknologisk lexikonmodul har medfört ett gynnsamt utgångsläge för ett gemensamt skandinaviskt samarbetsprojekt *Språkteknologiska lexikon för skandinaviska språk*, med akronymen *SkanLex-projektet.* Projektets övergripande målsättning är att skapa förutsättningar för sammanlänkningen av lexikonmodulerna för danska, norska och svenska och därmed lägga grunden till en flerspråkig lexikalisk databas.

Kännetecknande för den här aktuella modellen är att länkningen genomförs dels med stöd av den i SIMPLE-lexikonen införda informationen, dels med hjälp av en extern lexikonmodul som kommer att användas som ett gemensamt metaspråk, ett interlingua. En engelsk ordbok, *The New Oxford Dictionary of English* (1998), (NODE), kommer att prövas i denna funktion. För att ytterligare effektivisera arbetet med upprättandet av ekvivalensrelationer mellan källspråken och metaspråket kommer vi använda information i maskinläsbara tvåspråkiga lexikon mellan de enskilda skandinaviska språken och engelska.

Vid första ögonkastet kan tanken att länka samman ordbetydelser i de tre skandinaviska språken med hjälp av engelska som interlingua väcka förundran, men vid närmare eftertanke framgår interlingua-modellens överlägsenhet tydligt särskilt i jämförelse med transfermodellen. Medan antalet länkar i interlingua-modellen är lika med antalet källspråk, blir antalet länkar som måste etableras i transfer-baserade modellen långt större eftersom alla språkpar länkas direkt till varandra, dvs. till de enskilda målspråken (för n källspråk blir det n(n-1) länkar). Interlingua-modellen bidrar alltså till en avsevärd tids- och arbetsbesparing vid upprättandet av en flerspråkig lexikalisk databas. Dessutom kan, tack vare interlinguamodulen, flera språkteknologiska databaser kopplas samman genom sammanlänkning av enheter i deras respektive interlingua-moduler, och därmed kan databasernas innehåll och omfång mycket snabbt och effektivt breddas. Som ett exempel på en flerspråkig datatabas uppbyggd efter interlingua-principen kan nämnas EuroWordNet (Vossen 2001). Sammanlänkningen av EuroWordNet och SkanLex skulle automatiskt resultera i länkar till nya språk, som nederländska, spanska, italienska, tyska, franska, tjeckiska, estniska samt det engelska WordNet och dess nät av explicit markerade semantiska relationer (Fellbaum 1999).

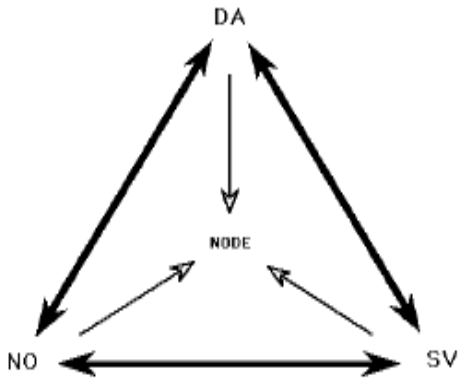
Arbetet med upprättandet av flerspråkiga lexikaliska databaser kan förväntas ge stöd åt såväl forskning inom traditionell och datamaskinell lexikologi som tillämpningar inom maskinöversättning, pre- eller posteditering av texter samt informationssökning. Informationen i dessa databaser kan också återanvändas för att generera pappersordböcker och elektroniska ordböcker avsedda för människor.

4.2 Huvuddragen i SkanLex-modellen

SkanLex-modellen vilar på tre huvudantaganden: 1) att ekvivalensrelationen är transitiv, vilket möjliggör sammanlänkning av ekvivalenter i de skandinaviska språken, 2) att användningen av en gemensam interlingua-modul effektiviserar sammanlänkningen och 3) att ju mer formaliserad semantisk information som finns att tillgå i de respektive språkteknologiska lexikonen, desto fler ekvivalenter kan automatiskt länkas samman. Antagandet om ekvivalensrelationens transitivitet innebär att om en ordbetydelse i något av de skandinaviska språken är ekvivalent med en ordbetydelse i NODE och denna i sin tur är

ekvivalent med ordbetydelsen i ett annat skandinaviskt språk så kan betydelserna i de båda skandinaviska språken betraktas som ekvivalenta. Med stöd av denna ekvivalensprincip kan länkningen mellan de skandinaviska språken utföras automatiskt för lexikonenheter som är fullekvivalenta, under förutsättning att länkningen till NODE har utförts. Även stora delar av partiellt ekvivalenta par kan förmodligen länkas samman automatiskt, men länkningen av dessa kräver ofta mer komplicerade länkningsprocedurer som bygger på access till ytterligare semantisk information om t.ex. deras ontologiska typ, domän, hyperonym, argumentstruktur eller argumentens selektionsrestriktioner. För att systematiskt kunna spåra betydelseskillnader hos partiellt ekvivalenta par specificeras deras typer och variationen dokumenteras.

Sammanlänkningsprocessen innefattar två delfaser, som man kunde kalla *metalänkning* och *skanlänkning. Metalänkning* avser parlänkning av lexem, mellan de nordiska språken å ena sidan och NODE å den andra sidan. Den länkningen utförs dels manuellt, dels automatiskt med hjälp av maskinläsbara tvåspråkiga ordböcker. (Se 4.3 nedan för en mer utförlig beskrivning.) Därefter kommer *skanlänkningen* som, med utgångspunkt i den information som tillkommit till följd av metalänkningen, etablerar länkar mellan lexemen i de skandinaviska språken. Dessa länkar skapas till största delen automatiskt. De två länkningsfaserna kan visualiseras på följande sätt:



Som exempel på metalänkningen kan vi ta det svenska ordet *kastanj* som i *Norstedts svenska ordbok* beskrivs på följande sätt

kastan'j subst. *kastanjen kastanjer*
1 typ av stor, oregelbunden nöt med (röd)brun färg varav vissa sorter är ätliga; urspr. omgiven av mjukt, taggigt ytterskal: *kastanj(e)puré; rostade kastanjer* □ äv. om (de arter av) större lövträd som bär denna frukt: *kastanj(e)allé; hästkastanj; äkta kastanj* □ i sms. äv. för att ange rödbrun färg: *kastanj(e)rött hår *kratsa/raka kastanjerna ur elden* (*för ngn*) hjälpa (ngn) ur en svår situation ofta genom egen risktagning 2 hård, hornartad utväxt på insidan av hästens ben.

Följande länkar etableras till *chestnut* i NODE:

kastanj 1/1/0	[Fruit]	*chestnut* 1/1/0
kastanj 1/1/1	[Plant]	*chestnut* 1/2/0
kastanj 1/1/2	[Colour]	*chestnut* 1/1/1
kastanj 1/2/0	[Organic_object]	*chestnut* 1/1/3

I sammanställningen ovan har vi inkluderat ontologisk information, hämtad ur det svenska SIMPLE-lexikonet, vilket åskådliggör ordets polysemi. Lemma-, kärn- och underbetydelsemarkeringar ovan följer dels den modell som tillämpats i Göteborgs lexikaliska databas, och i ordböcker som genererats ur denna databas, bl.a. *Norstedts svenska ordbok*, dels den som förekommer i NODE. Den första siffran anger lemma, den andra kärnbetydelse och den tredje, med undantag av nollan, hänvisar till underbetydelsen. All denna information, tillsammans med nyanserade betydelsebeskrivningar, gör att länkarna i metalänkningsfasen håller en hög kvalité, vilket garanterar korrekta länkningar även mellan de skandinaviska språken. Det kan noteras att vi kommer att lägga till nya betydelser eller betydelsenyanser till dem som redan finns i vår interlingua-modul, NODE, särskilt när sådana tillägg motiveras av förde skandinaviska språken karakteristiska begrepp som t.ex. *älv, syskon, mormor* eller om de av andra skäl saknas i NODE.

4.3 Metalänkningsstrategier i fokus

I det följande ger vi en kort översikt över hur metalänkningen kan utföras med stöd av information i lexikaliska databaser och maskinläsbara lexikon. Vi utgår från följande antaganden: (i) att metalänkningen kan genomföras först efter att mångtydighet hos lexikonenheter i källspråken och deras ekvivalenter har upplösts, deras betydelser fastställts samt graden av ekvivalens kartlagts och (ii) att metalänkningen av lexikonenheter som uppfyller ekvivalenskravet anpassas efter typen av mångtydighet som lexikonenheter i källspråket och deras engelska ekvivalenter uppvisar. Det faktum att den genomsnittliga graden av polysemi i svenska ligger på 1,6 för uppslagsord i Göteborgs lexikaliska databas (Malmgren 1988) och att liknande siffror noteras för engelska ger en ytterligare tyngd åt vikten att fokusera polysemifaktorn vid etableringen av ekvivalensrelationer.

Med tanke på att det rör sig om hantering av stora mängder av lexikaliska data efterlyses här lösningar och strategier som antingen leder till att länkningen kan utföras automatiskt eller som tillför ett maximum av stöd åt den manuella länkningen. Sådana lösningar blir lättare att åstadkomma om man tar hänsyn till och anpassar länkningsprocedurer efter den polysemifaktor lexikonenheter i käll- och metaspråket uppvisar. För att utforska detta påstående undersöker vi nedan följande huvudtyper av ekvivalensrelationer uppställda efter polysemifaktorn hos lexikonenhet i källspråket och dess engelska ekvivalent: (a) en monosem lexikonenhet i källspråket och en monosem ekvivalent, (b) en monosem lexikonenhet i källspråket och en polysem ekvivalent, (c) en polysem lexikonenhet i källspråket och en blandning av monosema och/eller polysema ekvivalenter.

Som tidigare nämnts inriktar vi oss främst på typer av ekvivalenspar med en likvärdig semantisk innebörd, vilket innebär att ekvivalensrelationer med semantiskt underspecificerade ekvivalenter (som i fallet *faster* aunt*, moster* aunt) eller överspecificerade, (som i *tak* ceiling, *tak* roof) samt andra par som uppvisar bristande ekvivalens ej tas upp i denna översikt

i och med att de fordrar ett mera individualiserat betraktelsesätt och därför i mindre grad berörs av de nedan föreslagna metalänkningstrategierna. (För en översikt över typer av ekvivalensrelationer representerade i tvåspråkiga ordböcker hänvisas till Svensén 1997:134-157).

(a) En monosem lexikonenhet i källspråket och en monosem ekvivalent

Monosema källspråksenheter som svarar mot monosema ekvivalenter hör till de lexikonenheter som automatiskt kan sammanlänkas med relativt enkla metoder eftersom de inte berörs av disambigueringsproblematiken. Listor över monosema ord i de berörda språken kan extraheras ur respektive språkens lexikaliska databaser. Dessa listor kan vidare samköras med hjälp av maskinläsbara tvåspråkiga lexikon, vilket bör resultera i framtagningen av två typer av ekvivalentenheter. Den ena typen är ekvivalentenheter med två element, där det råder ett-till-ett-relation mellan källspråksenheten och dess ekvivalent, vilket exemplifieras nedan:

visitkort	visiting-card
vitpeppar	white pepper
lyncha	lynch

Den andra typen är ekvivalentenheter som uppvisar ett-till-flera-relation, där en källspråksenhet svarar mot flera monosema ekvivalenter vilka står i närsynonymirelation till varandra, som i:

blindtarm	blind gut, caecum
primula	primula, primrose

Informationen om vilka och hur många ekvivalenter som finns utgör en av flera parametrar i deras flerdimensionella beskrivning i den flerspråkiga databasen, där fokus ligger på en kvalitativ beskrivning av ekvivalensrelationer. För att kunna dels explicitgöra eventuella skillnader mellan källspråksenheter och deras engelska ekvivalentvarianter, dels fånga deras gemensamma egenskaper kommer vi att i vår databas utgå från ekvivalentpar som basenheter vilka består av en källspråksenhet och en målspråksenhet (Jerker Järborg personlig kommunikation). Detta medför att ekvivalentenheter av den ovannämnda typen kommer att splittras i flera ekvivalentpar. För exemplet ovan etableras följande ekvivalentpar: {*blindtarm* blind gut}, {*blindtarm* caecum} och {*primula* primula}, {*primula* primrose}, vilket bl.a. möjliggör att i databasen fånga betydelseskillnader gällande likheter och skillnader, likheter i deras ontologiska typ och skillnader i deras domän och bruksvärde m.fl.

(b) En monosem lexikonenhet i källspråket och en polysem ekvivalent

Metalänkningen av monosema lexikonenheter i källspråket till polysema ekvivalenter i metaspråket förutsätter tillgång till på förhand disambiguerade ekvivalenter för att man skall kunna etablera korrekta ekvivalentpar. Information om ekvivalenters ordbetydelser finns att hämta i NODE, vår interlingua-modul. Informationen där är stringent och informationsrik, dock ej formaliserad enligt SIMPLE:s ontologiska modell, vilket försvårar den automatiska

sammanlänkningen av ordbetydelser i dessa två resurser. För att bemästra detta problem kommer vi att försöka att successivt överföra informationen från våra SIMPLE-lexikon till ekvivalenter i tvåspråkiga lexikon och vidare till lämpliga betydelsedistinktioner i NODE. Låt oss illustrera detta med ett exempel, ett monosemt substantiv *saffran* i svenska vars ekvivalent i engelska är det polysema *saffron* som står för både en krydda-betydelse och en växt-betydelse.

Sve-SIMPLE	Ontologi	Sve-eng	NODE
saffran 1/1/0	[Flavouring]	saffron	saffron 1/1/0 = [Flavouring] saffron 1/2/0 = [Plant]

Den ontologiska informationen om att ordet *saffran* tillhör klassen Flavouring i SIMPLE-ontologi överförs till dess ekvivalent *saffron* i det svensk-engelska lexikonet, för en vidare överföring till en lämplig betydelsedistinktion i NODE, vilket i detta fallet är *saffron* 1/1/0. Denna länkning, i kontrast till den föregående, utförs här manuellt eftersom det saknas information som kunde ha en disambiguerande effekt. Som exempel på en sådan informationstyp som kunde bistå den automatiska metalänkningen kan nämnas informationen om domäner, som i något olika former presenteras i dessa tre resurser. För att bättre kunna återanvända denna domäninformation bör graden av överlappning mellan domänmarkörer i de involverade resurserna kartläggas samt översättningsalgoritmer formuleras. Skulle det finnas domänuppgifter för *saffron* i NODE tillgängliga kunde man metalänka *saffran* 1/1/0 med *saffron* 1/1/0 automatiskt och därmed lämna *saffron* 1/2/0 olänkad vid det aktuella tillfället. Denna Plant-betydelse aktiveras först som ekvivalent då metalänkningen av *saffranskrokus* i svenska kommer i fråga.

(c) *En polysem lexikonenhet i källspråket och en blandning av monosema och/eller polysema ekvivalenter*

Den automatiska metalänkningen av polysema lexikonenheter är mer komplicerad i jämförelse med monosema på grund av att flera länkar mellan källspråksenheter och deras engelska ekvivalenter måste upprättas samt att flera parametrar kan behövas för att på ett distinkt sätt kunna disambiguera och beskriva dessa lexikonenheters polysema ordbetydelser. Till de högprioriterade parametrar som antas specificera ett ekvivalentpar räknas lexikonenheters ontologiska typ, domän och hyperonym. Något lägre prioritet ges åt informationen om ords kontextuella egenskaper sådana som argumentstruktur och selektionsrestriktioner samt stilnivå och bruksvärde. Metalänkningen av polysema lexikonenheter följer i huvuddrag samma tillvägagångssätt som tillämpas för lexikonenheter av typ (b) ovan, med undantag för att de enskilda länkningsprocedurerna upprepas tills alla betydelsedistinktioner i källspråket antingen bildar ett ekvivalentpar med en befintlig lexikonenhet från NODE eller resulterar i ett nytillskott till NODE:s ursprungliga uppsättning av lexikonenheter.

De olika moment som ingår i en metalänkning beskrivs nedan med utgångspunkt i det som gäller för det polysema ordet *kalv*. De består av följande delmoment: (i) den ontologiska informationen om ordet *kalv* hämtas ur SIMPLE-lexikonet, (ii) ordet slås upp i ett svensk-engelskt lexikon och ekvivalenterna *calf* och *veal* samt tillhörande information hämtas,

(iii) *calf* och *veal* slås upp i NODE för att undersöka deras polysemifaktor och betydelsedistinktioner innan ekvivalentparen kan upprättas.

Sve-SIMPLE	Ontologi	Sve-eng	NODE
kalv 1/1/0	[Animal]	calf	calf 1/1/0 (+1/1/1)
kalv 1/1/1	[Food]	veal	veal 1/1/0
kalv 1/1/2	[Physical object]	calf 2/1/0	
kalv 1/1/3	[Material]	calf	calf 1/1/2
(kortform för kalvskinn)			

I det första delmomentet inkluderas även automatiska subrutiner som undersöker om den ontologiska informationen är betydelseskiljande; om så inte är fallet, kompletteras den med information om domäner och det kontrolleras på nytt om den utökade uppsättningen av information har en disambiguerande effekt. Proceduren upprepas med nya semantiska parametrar tills det polysema ordet blir disambiguerat.

I det andra delmomentet återanvänds informationen om ordets engelska ekvivalenter och deras betydelseförklaringar ur Norstedts tvåspråkiga svensk-engelska ordbok. De i artikeln infogade betydelseförklaringarna kan ofta med hjälp av översättningsalgoritmer relateras till SIMPLE:s ontologiska kategorier. Exempelvis kan förklaringar åsyftande djurbetydelse likställas med Animal och kött-betydelse med Food, vilket bistår den automatiska sammanlänkningen av enheter i SIMPLE-lexikonet med deras engelska ekvivalenter. Kalvskinn-betydelse är också en möjlig kandidat för en automatiskt sammanlänkning på basis av dess ontologiska klass. Däremot måste betydelsen av *kalv* 1/1/2 [Physical object], "ngt mindre (bihang) i förh. till ngt större, spec. i namn på udde, mindre ö utanför större e.d.g.", översättas med utgångspunkt i dess beskrivning i SIMPLE-lexikonet och Göteborgs lexikaliska databas, eftersom den varken förekommer i den tvåspråkiga ordboken eller har sin ekvivalent i NODE.

I det tredje delmomenet kan enbart metalänkningen för ekvivalentparet {kalv, veal} genomföras fullt ut automatiskt i och med att dess ekvivalent är monosem. Kalv-betydelser för [Animal], [Material] länkas till NODE:s betydelsedistinktioner manuellt. Manuellt läggs också en nytillskottsekvivalent "calf holme" till NODE som svarar mot *kalv* 1/1/2 [Physical object]. Det bör noteras att den betydelse som ges i NODE för *calf* 2/1/0 "a floating piece of ice detached from an iceberg" som egentligen också tillhör till [Physical object]-kategorin ej är ekvivalent med *kalv* 1/1/2. Detta visar att de ekvivalentpar som har sammanlänkats enbart på basis av ontologiska kriterier bör efterkontrolleras med hjälp av flera semantiska parametrar för att verifiera deras ekvivalensstatus.

5. Slutord

I vår vision av en flerspråkig databas för skandinaviska språken ryms även tre tvåspråkiga databaser där de enskilda skandinaviska språken sammanlänkas till engelska. Denna vision kännetecknas av både bredd och djup. Bredden visar sig främst i den flerspråkiga databasens potential att på ett smidigt sätt integereras med andra lexikaliska databaser via en interlingua-modul och därmed kunna öka dess innehåll både när det gäller nya språk och informationsvidden. Djupet blir tydligt då de semantiska dimensionerna fångas med hjälp av parametrar som lexikonenhetens betydelse, dess polysemifaktor, ontologiska klass, do-

män, hyperonym, argumentstruktur, selektionsrestriktioner, synonymi, logiska polysemi, stilvärde, bruksvärde, vilka inverkar på upprättandet av ekvivalentpar och beskrivningen av deras interna relationer. Genom att skapa förutsättningar för att upprätta en flerspråkig databas bäddar man för framtagning av både standardlösningar för formalisering av lexikaliska data och för utarbetande av strategier för återanvändning av lexikalisk information från olika maskinläsbara lexikaliska resurser.

[1] Eksempler på ontologiske typer som de enkelte ordbetydninger tilskrives er f.eks. *Instrument* ('gaffel'), *Cognitive Fact* ('viden') *Relational State* ('forbindelse'), *Animal* ('giraf') , *Profession* ('amtsborgmester') etc.

[2] Senseval er et internationalt projekt der omhandler betydningsopmærkning af korpuseksempler med det formål at opbygge trænings- og evalueringsmateriale til computerprogrammer der skal kunne foretage automatisk betydningsbestemmelse af flertydige ord (Kilgarriff 1998).

Referencer

Faarlund et al. 1996. *Norsk referansegrammatikk.* Oslo.

Fellbaum, Ch. 1999. Semantics via Conceptual and Lexical relations. I E. Viegas, ed. *Breath and Depth of Semantic Lexicons*. Dordrecht: Kluwer Academic Publishers.

Kilgarrif, A. 1998. SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. In Proceedings from First International Conference on Language Resources and Evaluation pp.581-588, Granada, Spanien.

Lenci, A. F. Busa, N. Ruimy, E. Gola, M. Monachini, N. Calzolari, A. Zampolli, J. Pustejovsky, E. Guimier, G. Recourcé, L. Humphreys, U. Von Rekovsky, A. Ogonowski, C. McCauley, W. Peters, I. Peters, R. Gaizauskas, M. Villegas, O. Norling-Christensen. 2000a. *SIMPLE Linguistic Specifications*, Technical Report, University of Pisa.

Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters. W. Peters, N. Ruimy, M. Villegas, A. Zampolli. " SIMPLE – A General Framework for the development of Multilingual Lexicons", in: T. Fontenelle (ed.) *International Journal of Lexicography Vol 13*. pp. 249-263.2000. Oxford University Press.

Levin, Beth 1993. *English verb classes and alternations.* Chicago.

Malmgren, S-G. 1988. On Regular Polysemy in Swedish. I *Studies in Computer-Aided Lexicology*. Stockholm: Almquist & Wiksell International.

Nordgaard, Torbjørn: NorKompLeks 1996. Some Linguistic Specifications and Applications, 1996. In Lindebjerg, Ore & Reigem: *ALLC-ACH '96. Abstracts.* Humanistisk Datasenter, Universitetet i Bergen.

Norstedts svenska ordbok. 1999. I *Ordboken*, version 2.0.5 utarbetad av C. Wihlborg och J. Engstrand. Stockholm: Datadromeda.

Pedersen, B.S. & S. Nimb 2000. Semantic Encoding of Danish Verbs in SIMPLE - Adapting a verb-framed model to a satellite-framed language. *Proceeding from Second International Conference on Language Resources and Evaluation, LREC 2000*, Athens.

Pedersen, B.S., R. V. Fjeld, M. Toporowska Gronostaj 2002. Harmonisering og sammenkædning af sprogteknologiske ordbaser med særligt henblik på informationssøgning- en rapport fra SPINN-netværket, NorFa Årsskrift 2001.

Pedersen, B.S., R.V. Fjeld, M. Toporowska Gronostaj 2001. SPINN: SPråkteknologi och INformationssökning i Norden. I *LexicoNordica* 8. Oslo.

Ruimy, N. O. Corazzari, E. Gola, A. Spanu, N. Calzolari, A. Zampolli (1998). 'The European LE-PAROLE Project: The Italian Syntactic Lexicon', in: *First International Conference on Language Resources & Evaluation*, Granada, Spain.

Runde, Ålov 2000. Korpusoppbygging ved Seksjon for leksikografi og målføregransking. *Ord om ord 6, Årsskrift for leksikografi 2000*, Seksjon for leksikografi og målføregransking, Oslo.

Sveen, Andreas 1996. *Norwegian Impersonal Actives and the Unaccusative Hypotesis.* Oslo.

Svensén, B. 1987. *Handbok i lexikografi*. Stockholm: Esselte Studium.

The New Oxford Dictionary of English (NODE). 1998. Ed. Judy Pearsall. Oxford: Oxford University Press

Vossen, P. 2001. Condensed Meaning in EuroWordNet. I Bouillion, P. och F. Busa (eds.), *The Language of Word Meaning*. Cambridge:University Press.