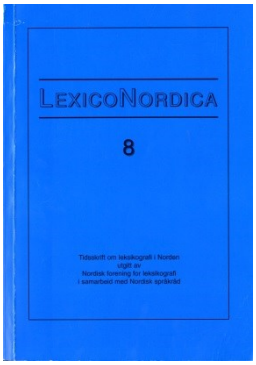
LexicoNordica

Titel:	SPINN: SPråkteknologi och INformationssökning i Norden
Forfatter:	Bolette Pedersen, Ruth Vatvedt Fjeld og Maria Toporowska Gronostaj
Kilde:	LexicoNordica 8, 2001, s. 125-138
URL:	http://ojs.statsbiblioteket.dk/index.php/lexn/issue/archive



Søgbarhed
Artiklerne i de ældre LexicoNordica (1-16) 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 Vatvedt Fjeld & Maria Toporowska Gronostaj

SPINN: SPråkteknologi och INformationssökning i Norden

This paper describes a coordinate research project with the purpose of investigating the possibilities of a multilingual computational lexicon covering the Scandinavian languages. The initiative originates from the work carried out within the EU-project, SIMPLE, in which Danmark and Sweden participated, and which resulted in lexicon modules for 12 EU-languages including Danish and Swedish. The Danish lexicon is being reused for creating a parallel Norwegian module (see section 2). In the final section a proposal for linking the word meanings in these three lexicons by means of English as a common interlingua is sketched out. The SPINN network is the forum for cooperation among researchers in the Nordic countries on the issues concerning the use of computational lexicons for natural language processing purposes with specific focus on information retrieval.

1. Bakgrund och syfte

1.1. Inledning

Den övergripande målsättningen för forskarnätverket SPINN (SPråkteknologi och INformationssökning i Norden) är att bidra till att de forskningsmiljöer i Norden som arbetar med att bygga upp språkteknologiska lexikon för innehållsbaserad informationssökning samt de forskningsmiljöer som arbetar med innehållsbaserade sökmotorer, utbyter erfarenheter och inleder ett samarbete. Nätverket finansieras av NORFA (Nordisk ForskerAkademi) under en tvåårsperiod fr.o.m. januari 2001.

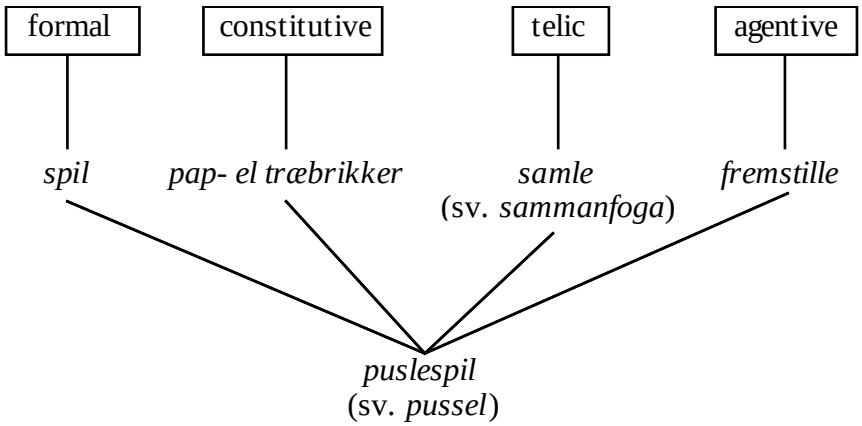
1.2. Språkteknologiska ordböcker för de nordiska språken

Det primära incitamentet till att SPINN-nätverket startades var det dansk-svenska samarbete som inleddes inom ordboksprojektet SIMPLE (Semantic Information for Plurilingual, Multifunctional Lexica, se Lenci et al. 2000). I detta EU-projekt har man skapat förutsättningar för att utarbeta språkteknologiska lexikon för 12 språk med vardera 10.000 ordbetydelser: 7000 substantiv, 2000 verb och 1000

adjektiv. För var och en av dessa ordbetydelser kan man tillföra information av typen (i) begreppstyp (semantisk klass), (ii) domän, (iii) definition, (iv) korpusexempel, (v) argumentstruktur (semantisk valens), (vi) selektionsrestriktioner, (vii) semantiska relationer och semantiska särdrag, samt slutligen synonymi-, polysemi- och kollokationsrelationer.

Ett av de grundläggande antaganden som görs i SIMPLE-modellen är att ords betydelse kan variera mycket med avseende på intern komplexitet och att denna variation kan beskrivas med en flerdimensionell ontologi (Lenci *et al.* 2000). Vissa ords betydelse kan beskrivas med hjälp av s.k. *simple* typer där man härleder information från en enda modernod i ontologin, medan andra är mer komplexa (*unified types*) eftersom de hämtar information från flera modernoder i ontologin. Dessa betydelsedimensioner uttrycks i SIMPLE med hjälp av en utvidgad s.k. Qualiateori som bygger på Pustejovsky (1995).

Om vi betraktar ett ord som det danska *puslespil* (sv. *pussel*) kan vi se på vilket sätt detta ords betydelse kan uttryckas med hjälp av de fyra Qualiarollerna.[1] Om vi slår upp ordet i en traditionell ordbok, t.ex. *Nudansk Ordbog*, finner vi följande definition av ordet: 'et spil med træ- eller papbrikker i forskellige faconer som lægges sammen så de danner et hele'. I SIMPLE-modellen uttrycks samma betydelsekomponenter som i den klassiska definitionen, men de har formaliserats för att passa in i den fyrdimensionella strukturen:



Det finns alltså fyra betydelsedimensioner för ordet *puslespil*: (i) den formella rollen (formal role), som ger information om ordets placering i

[1] För en närmare redogörelse för de danska och svenska SIMPLE-ordböckerna se Pedersen & Keson (1999), Pedersen & Nimb (2000), Nimb & Pedersen (2000), Pedersen (opublicerad) samt Kokkinakis, Toporowska Gronostaj & Warmenius (2000).

ontologin med hjälp av en *is-a*-relation (vilket motsvarar genusdelen av definitionen): ett pussel är ett slags spel; (ii) den konstitutiva rollen (constitutive role), som uttrycker ett brett spektrum av semantiska relationer som gäller ordets interna struktur (i detta fall *part-of*: att det består av ett antal bitar), (iii) den teliska rollen (telic role), som beskriver objektens typiska funktion (här en *used-for*-relation: ett pussel ska sammanfogas), och slutligen (iv) den agentiva rollen (agentive role), som beskriver objektens ursprung och som primärt gäller om objekten är frambringade av naturen eller av människohand (i detta fall en *made-by*-relation).

Den egentliga ordboksingången för *puslespil* framgår av nedanstående uppställning. I detta fall anges dessutom att det existerar en systematisk polysemirelation till en metaforisk användning av ordet *puslespil*, som i *det var et puslespil at stable udstillingen på benene*. Ordet har å andra sidan inte någon synonym, och således är detta fält i ordboksingången inte ifyllt.

Semantic Unit	*puslespil-ART* (artifact reading)
Definition	*et spil med træ- el. papbrikker i forskellige faconer som skal lægges sammen så de danner et hele* (NDO)
Corpus example	*nu var hun næsten ved at være færdig med det puslespil, hun var begyndt på lige efter påske*
Semantic type	Artifact
Unification path	Concrete-Entity|Agentive|Telic
Domain	General
Semantic class	Artifact
Formal quale	is-a = *spil*
Agentive quale	created-by = *fremstille*
Telic quale	used-for = *samle*
Constitutive quale	has-as-parts=*brikker*
Systematic polysemi	ArtifactAbstract= *puslespil-ABS*
Synonymy	nil

Det faktum att alla språkgrupperna i SIMPLE-projektet har använt samma specifikationer för beskrivningen gör att det är lättare att få till stånd en sammanlänkning av samtliga språk i framtiden. För att försäkra sig om ett gemensamt kärnordförråd i projektet, har man dessutom valt ut 1000 kärnbegrepp som för danska och svenska har utkristalliserat sig ur ca 1300 ordbetydelser. Utöver detta gemensamma ordförråd känner vi för närvarande inte till hur stor den samlade mängden betydelser i de två ordböckerna är, men vi tror att ungefär hälften av betydelserna (ca 5000) finns beskrivna i båda ordböckerna, eftersom ordurvalet är baserat på frekvenskriterier i båda språken. I avsnitt 3 beskrivs närmare

vilken strategi vi planerar att utprova i nätverket för att kunna koppla samman ordböckerna.

Samarbetet mellan Danmark och Sverige har efter hand utvidgats till att också omfatta en norsk samarbetspartner, då man vid Universitetet i Oslo ville skapa en norsk SIMPLE-ordbok med utgångspunkt från den danska. Detta arbete beskrivs ytterligare i avsnitt 2. Dessutom finns planer på att i framtiden också inkludera isländska i projektet – om än möjligen i något mindre skala.

1.3. Det teknologiska incitamentet för nätverket

Den starkt ökande användningen av Internet och tillgången till stora cd-romdatabaser ökar behovet av informationssökningssystem som är mer intelligenta och som innehåller mer språkligt vetande än dagens sökmaskiner. Innehållsbaserad informationssökning har således blivit ett betydelsefullt begrepp inom detta område och då avses sökning som grundar sig på en hög grad av begreppsmässigt och lexikaliskt vetande.

De nordiska länderna har, sett i ett europeiskt perspektiv, på flera områden varit föregångare inom den teknologiska utvecklingen. Under de senaste åren har det emellertid skett en massiv utveckling inom det språkteknologiska området när det gäller de större europeiska språken, och detta har bl.a. inneburit att långt mer avancerade verktyg för informationsbehandling har utvecklats för dessa. Denna utveckling innebär paradoxalt nog att de nordiska länderna får svårare och svårare att göra sig gällande i nätverkssamfundet – i varje fall på sina egna språk. Utvecklingen kan inte bara leda till att textmaterial på de nordiska språken till en viss grad blir "ointressanta" i informationsteknologiskt hänseende när de språkteknologiska verktygen inte kan hantera våra språk, utan den kan också få till konsekvens att de nordiska språken på längre sikt upphör att fungera som arbetsspråk och således inte längre utvecklas, t.ex. inom nytillkomna fackspråksdomäner.

Sett i detta perspektiv är det viktigt med ett gemensamt nordiskt projekt med sikte på att språkteknologiska resurser och verktyg för de nordiska språken kan utvecklas och följa med i den snabba utveckling vi kommer att få se inom detta område under de närmaste åren.

2. En norsk version av SIMPLE med utgångspunkt från den danska

2.1. Varför en norsk SIMPLE-ordbok?

Den uppgift inom lexikografin i Norge som under efterkrigstiden haft högst prioritet har varit dokumentation och kodifiering av det norska folkspråket, de norska dialekterna och användningen av den nya norska skriftspråksstandarden. Dessutom har språkpolitisk konkurrens medverkat till att projekt där den lexikografiska grundforskningen tagit sin utgångspunkt i moderna lingvistiska teorier har gynnats. Det har således bara gjorts spridda försök till formaliserad betydelsebeskrivning av det norska ordförrådet. I Danmark, Finland och Sverige har man under det senaste decenniet inom olika projekt arbetat med formaliserade maskinläsbara lexikon, medan Norge har halkat efter inom detta område. Därför är det oerhört viktigt för norsk lexikografi att delta i ett nordiskt nätverk för sammanlänkning och harmonisering av språkteknologiska ordböcker.

Genom *Dokumentasjonsprosjektet* och Textlaboratoriet vid Det historisk-filosofiske fakultet vid Universitetet i Oslo har en lexikalisk databas omfattande morfologin för ett lemmaurval som i huvudsak motsvarar *Bokmålsordboka* utarbetats. Vidare har inom NorKompLeks-projektet vid Norges teknisk-naturvetenskaplige fakultet i Trondheim en databas med formaliserad beskrivning av argumentstrukturen för verben i *Bokmålsordboka* skapats utifrån de exempelmeningar som finns i denna. Dessa arbeten är varken fullständiga eller enhetliga, så det är en stor utmaning att åstadkomma ett systematiskt språkteknologiskt lexikon för norska, så som vi planerar att göra det i det vi kallar *Leksikalsk database for moderne bokmål* (LDB-prosjektet).

Danska är det skandinaviska språk som ligger närmast norska. De danska lexikograferna har varit generösa nog att låta oss få tillgång till sin SIMPLE-bas, så att vi i stor utsträckning kan översätta och anpassa dess definitioner till norska. Därutöver sätter vi in definitionerna från *Bokmålsordboka* och väljer ut norska exempel ur norska textkorpusar. Detta arbete är vi i full gång med, men det är självfallet inte helt oproblematiskt.

I första omgången har vi försökt att finna norska ekvivalenter till de lemman som finns i den danska SIMPLE-basen och att föra in norska exempelmeningar, och som regel går detta bra. Här presenteras dock en del fall då det är problematiskt att finna norska ekvivalenter. De danska definitionerna kvarstår, och de norska har hämtats från *Bokmålsordboka*. Alla norska tillägg har tills vidare lagts in i en kopia av den danska basen, med bokstaven *N* tillagd som identifikation.

2.2. Full ekvivalens

I de allra flesta fall är ekvivalensen heltäckande, både formellt och betydelsemässigt:

```
naming="bar"
namingN="bar"
exampleN="Denne svensken tok SM-finalen i bar
 overkropp!"
freedefinition="som ikke er dækket af noget
 (NDONY)"
freedefinitionN="naken, udekket"

naming="lad"
namingN="lat"
exampleN="Sånn sett er pistolskyting kanskje en
 lat manns sport"
freedefinition="doven (NDONY)"
freedefinitionN="uvillig til å gjøre noe, uten
 energi og tiltakslyst"
```

När skillnaden mellan danska och norska bara består i systematiska ortografiska och morfologiska olikheter men betydelsen av orden är den samma räknas detta som full ekvivalens.

2.3. Ungefärlig ekvivalens

2.3.1. Olika form, samma betydelse

I en del fall finns inte det danska ordet i norskan, men det finns ett annat uttryck som motsvarar det helt eller tämligen bra:

```
naming="skrap"
namingN="streng"
freedefinition="som   stiller   store   krav   til
andres
  opførsel og indsats (NDONY)
freedefinitionN="hard, nådeløs"
```

Ordet *skrap* er enligt *Dansk Etymologisk ordbok* sannolikt lånat från det holländska *schrap*, men den formen används inte i norskan, vare sig med den generella betydelsen 'rask, kraftig' eller den mer speciella betydelsen 'streng', som är definierad här. Den norska ekvivalenten måste därför bli *streng* i betydelsen 'hard, nådeløs', som är delbetydelse 2 i *Bokmålsordbokas* definition av detta lemma. Detta är egentligen en metaforisk användning av betydelse 1 'barsk, hard'. Ett sådant ekvivalensförhållande ger inga problem i databasen.

2.3.2. Norsk ekvivalent med annan avledningshistoria

Adjektivet *rojal* har traditionellt inte funnits i norskan. I neutral betydelse har man använt *kongelig*. En person som är anhängare av kungamakten heter på norska *rojalist* och alltså måste en adjektiv-avledning av detta bli *rojalistisk*:

```
naming="royal"
namingN="rojalistisk"
freedefinition="som er tilhænger af kongedømme,
 el. som beskæftiger sig med kongelige personer
 (NDONY)"
freedefinitionN="kongeligsinnet"
```

Rojal är en form som är helt ny i norskan, kanske bara tre–fyra år gammal:

> Det hersker også betydelig **rojal** familieidyll om dagen. (*Dagbladet* 2001-03-29)

Att danskan har en grundform som adjektiv medan norskan har en avledning bör vara en tillfällighet. Det ger i alla fall inga problem i den norska databasen.

2.4. Avsaknad av ekvivalens

2.4.1. Total avsaknad av ekvivalens

Några danska ord finns överhuvudtaget inte i norskan:

```
naming="skåret"
freedefinition="som der er slået skår i(NDONY)"
```

På norska föreligger här ett lexikalisk lucka. Vi har substantivet *skår* i samma betydelse som i danskan, men kan inte konstruera ett adjektiv av formen i den betydelsen. I dialekterna finns former som *skalete, skalut, skælut, skælete*, möjligen bildade av det fornnorska *skar›*, men de finns inte belagda i någon av de undersökta ordböckerna för standardbokmål, och det är därför inte möjligt att ta med dem i en databas över det centrala norska ordförrådet.

Ekvivalent saknas också för följande lemma:

```
naming="mellemøstlig"
```

```
freedefinition="som har at gøre med Mellemøsten
 (NDONY)"
```

På norska måste denna betydelse uttryckas med en prepositionsfras eller en sammansättning: *Midtøstenkonflikten, konflikten i Midtøsten.* Att norskan inte har *skåret,* kan bero på att denna betydelse delvis täcks av *hakkete,* och uttrycket *mellomøstlig* är upptaget av det som många nu efter engelskt mönster kallar *tverrøstlig. Mellom* har möjligen en vidare betydelse i danska än i norska. Sådana här förhållandevis små ekvivalentskillnader kan det vara besvärligt att upptäcka om man inte gör systematiska studier.

2.4.2. Avsaknad av formell ekvivalens

Samma betydelse finns på båda språken, men den realiseras formellt på olika sätt:

```
naming="sindet"
freedefinition="som har en bestemt indstilling
el.
  hensigt (NDONY)"
```

Detta lemma kan i modern norska normalt bara stå i sammansättningar som *vennligsinnet, ondsinnet, frisinnet, storsinnet, norsksinnet* och liknande. Detsamma gäller *flippet:*

```
naming="flippet"
freedefinition="som lever frit og uden særlige
 normer"
```

Norskan har här bara sammansatt form *utflippet*:

> Asbjørn har i løpet av få uker utviklet seg fra en pen forstadsjournalist i rutete bukser til en utflippet hippie. (*AA* 99-05-12)

Dessa danska former måste därför utgå i den norska databasen, eller eventuellt ingå som sammansättningar.

2.4.3. Avsaknad av delbetydelse

Några gånger har danskan delbetydelser av ett gemensamt ord som inte har någon ekvivalent i norskan. Ett exempel på detta är *sjofel:*

```
naming="sjofel"
namingN="uanstendig"
freedefinition="seksuelt anstødelig (NDONY)"
freedefinitionN="usømmelig"
```

Vi kan inte finna några belägg för att *sjofel* i norskan används om det som är sexuellt anstötligt, det vi skulle kalla *uanstendig*, och som väl helt motsvarar det danska *uanstændig*. Konsekvensen bör därför bli att denna betydelse av *sjofel* stryks i den norska databasen.

I grundbetydelsen kan ekvivalensen vara nästintill total, medan en betydelseutveckling som finns i det ena språket kan saknas i det andra. Detta gäller särskilt vid överförd betydelse:

```
naming="skæv"
namingN="skjev el. skeiv"
freedefinition="som ikke er lige men burde være
 det (NDONY)"
freedefinitionN="forkjært, vrang"
```

Så som detta adjektiv definieras i danskan har det ett modalt innehålls-element 'men burde være det'. I sin grundbetydelse är adjektivet ett rent relationellt dimensionsadjektiv, skevhet kan beräknas objektivt. I vilken grad en skevhet är acceptabel eller inte kan inte läsas ut av adjektivet i sig, men självfallet kan det stå i en kontext som tillför en sådan tolkning. När adjektivet används modalt och normativt har det en svag metaforisk betydelse och det kan vara svårt att avgöra när en metaforisk användning är så lexikaliserad att den bör finnas med som en del av en lexikalisk definition. *Bokmålsordboka* har som betydelse 2 'forkjært, vrang', och detta är en överförd, metaforisk betydelse jämfört med 'som ikke er lige men burde være det'. Det handlar då inte längre om en fysisk egenskap, och således är det ett annat lemma. Ordet används ofta i denna betydelse i modern norska: *skjev aldersfordeling*, *skjev kjønnsfordeling*, *skjev film*.

I norska finns också en variant med diftong, *skeiv*. I det undersökta bokmålsmaterialet har denna form alltid överförd betydelse:

> **Skeive** kalles de fordi homofil og lesbisk bevegelse står bak arrangementet (*Stavanger Aftenblad* 99-09-11)

> Sjøl ikke ivrige Bergens-journalister klarte å få mer enn **skjeive** (sic) blikk i garderoben (*Dagbladet* 99-09-12)

Denna specialisering motiverar att *skeiv* bör införas som ett eget lemma i norskt bokmål.

2.5. Tveksamt om norsk ekvivalent finns

I vissa fall är det svårt att dokumentera någon norsk ekvivalent till det danska ordet. Detta gäller särskilt när lemmat är slangbetonat eller inte särskilt frekvent i det man räknar som det centrala ordförrådet. Det måste diskuteras i vilken utsträckning sådana lemman ska tas med i den norska versionen av databasen:

```
 naming="slikket"
 freedefinition="  (neds.)  som  er  overdrevent
pæn,
  ordentlig og nydelig(NDONY)"
```

Norskan har givetvis den konkreta betydelsen av *slikket* som motsvarar participet av *å slikke*, men i den specialiserade betydelsen 'for pen' används inte detta ord. Ordet uttrycker en värdering och visar tydligt hur härledda betydelser kan vara mer språkspecifika än sina grundbetydelser. *Dansk Etymologisk Ordbog* är inte helt uppdaterad i sin beskrivning av norska här, då den under *slikket* 2 uppger: "(eft. ty. *geleckt*) 'pyntet, oversirlig'; no. d.s." I nyare norska används *glatt* i denna betydelse, men det är en något vidare karakteristik. I modern slang används *glætt* med ett främre a-ljud i samma betydelse som danskans *slikket*, men om sådana slanguttryck ska anses som allmänna nog är tveksamt. Man finner tre belägg i norskt tidningsmaterial, t.ex.:

> ABBA var for glætt (*Dagbladet* 99-04-06)

2.6. Sammanfattning

En del av ekvivalensproblemen uppstår därför att varje betydelse måste representeras av ett eget lemma, dvs. att lemmana inte kan ha polysemer. En sådan polysemiuppdelning är nödvändig om man ska kunna använda lexikonet exempelvis för automatisk översättning mellan danska och norska. Ett sådant tillvägagångssätt visar dock att det finns en avsaknad av ekvivalens mellan danska och norska och att anpassningen av det danska SIMPLE följaktligen kommer att bli långt mer omfattande än ett rent översättningsarbete .

De få exempel som här redovisats visar också på de problem vi måste vara förberedda på om vi genomför SkanLex-projektet där danska, svenska och norska ska kopplas samman i ett och samma lexikon med engelska som Interlingua (se avsnitt 3). Här har vi emel-

lertid koncentrerat oss på problematiken vid översättning av det danska lexikonet. I de allra flesta fallen är ekvivalensen total, något som betyder att tillgången till den danska SIMPLE-basen är en värdefull gåva till norsk lexikografi.

3. SkanLex-projektet

3.1 Projektets övergripande syfte och frågeställningar

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 skapat 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 sammanlänkning av lexikonmodulerna för danska, norska och svenska och därmed lägga grunden till en flerspråkig lexikalisk databas.

I de föregående avsnitten har vi berört problematiken kring formaliseringen av ordbetydelser samt etableringen av ekvivalensrelationer med utgångspunkt i danska och norska, vilket antyder vilken typ av teoretiska frågor som kommer att aktualiseras inom SkanLex-projektet. Eftersom den faktiska sammanlänkningen av ordbetydelser kan genomföras först efter att ordets mångtydighet har upplösts, dess betydelse(r) fastställts samt graden av ekvivalens kartlagts, kommer teorin och praktiken att flätas samman under arbetets gång. Därmed hoppas vi uppnå bästa möjliga resultat när det gäller anpassningen av länkningsstrategier till den stora mångfalden av ordbetydelser i våra lexikon.

Kännetecknande för den här aktuella modellen är att länkningen genomförs dels med stöd av informationen i SIMPLE-lexikonen, 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.

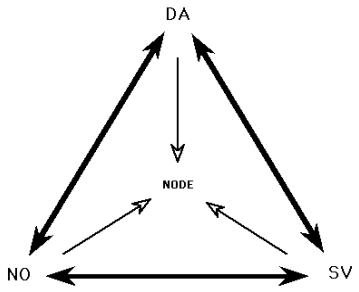
I det följande skisseras några huvudrag av den Interlingua-baserade modell som vi tänker tillämpa i SkanLex-projektet.

3.2 Huvuddragen i SkanLex-modellen

SkanLex-modellen bygger 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 helt ekvivalenta, under förutsättning att länkningen till NODE har utförts. Även stora delar av (partiellt ekvivalenta) par (tripletter etc.) kan förmodligen länkas samman automatiskt, men länkningen av dessa kräver ofta mer komplicerade procedurer som bygger på access till ytterligare semantisk information om t.ex. deras ontologiska typ, domän, hyperonymer, 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 kan kalla *metalänknin*g och *skanlänkning. Metalänknin*g 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. 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 metalänkning kan vi ta det svenska ordet *kastanj* som i *Svensk ordbok* beskrivs på följande sätt

> **kastan'j** [Norstedts svenska ordbok] 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 underbetydelsemarkeringarna ovan följer dels den modell som tillämpats i *Svensk ordbok* och motsvarande databas, 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 förbehåller oss rätten att tillföra nya betydelser eller betydelsenyanser till de till dem som redan finns i vår Interlingua-modul, NODE, särskilt när sådana tillägg motiveras av för skandinaviska språken karakteristiska begrepp som t.ex. *älv*, *syskon*, *mormor* eller om de av andra skäl saknas i NODE.

3.2 Slutord

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 en transfer-modell. Medan antalet länkar i Interlinguamodellen 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 flerspråkiga språkteknologiska lexikon.

Dessutom kan, tack vare Interlingua-modulen, flera språkteknologiska lexikon eller nätverk kopplas samman genom sammanlänkning av enheter i deras respektive Interlingua-moduler, och därmed kan lexikonets innehåll och omfång snabbt och effektivt breddas.

Arbetet med de språkteknologiska lexikonen kan förväntas ge stöd åt såväl forskningen inom traditionell och datamaskinell lexikologi, som tillämpningar inom maskinöversättning, pre- eller posteditering av texter samt informationssökning. Informationen i dessa lexikonmoduler kan också återanvändas för att generera pappersordböcker och elektroniska ordböcker avsedda för människor.

Litteratur

Fjeld, R.V. 1999: Leksikalisk database for moderne bokmål – LDB. I: *Ord om Ord* 5. Årsskrift for leksikografi. Oslo.

Kokkinakis D., M. Toporowska Gronostaj, & K. Warmenius 2000: Annotating, Disambiguationg & Automatically Extending the Coverage of the Swedish SIMPLE Lexicon. I: *Proceedings from the*

Second International Conference on Language Resources and Evaluation. Athens, 1397–1405.

Lenci, A., N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas & A. Zampolli. 2000: SIMPLE – A General Framework for the Development of Multilingual Lexicons. I: *International Journal of Lexicography*, Vol. 13, 249–263.

Nielsen, N. Å. 1976. *Dansk Etymologisk Ordbog*. København.

The New Oxford Dictionary of English (NODE). 1998. Oxford.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA.

Pedersen, B.S. & B. Keson 1999: SIMPLE – Semantic Information for Multifunctional Plurilingual Lexicons: Some Examples of Danish Concrete Nouns. I: *SIGLEX 99: Standardisaing Lexical Resources*. ACL Workshop, 1999. University of Maryland, USA, 46–51.

Pedersen, B.S. & S. Nimb. 2000: Semantic Encoding of Danish Verbs in SIMPLE – Adapting a verb-framed model to a satellite-framed language. I: *Proceedings from the second Internal Conference on Language Resources and Evaluation*. Athens, 1405–1412.

Pedersen, B.S. A Danish Semantic Lexicon and its Application in Content-based Information Retrieval. Opublicerad, utkommer i Bouillon & Viegas (eds.), French Journal T.A.L. *Semantic Lexicons in Natural Language Processing*. Hermes, France.

Svensk ordbok. Stockholm 1999. CD-version 2.0.5

Wangensteen, B. og M. Landrø 1993: *Bokmålsordboka*. Oslo.

Se även:
http://cst.ku.dk/projects/spinn/spinnhome.html
http://spraakdata.gu.se/simple/simple_parole-index.html

(Översättning av avsnitt 1 och 2 från danska resp. norska:
Maia Andréasson)