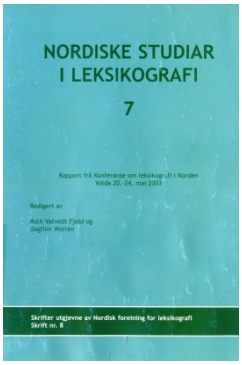
# NORDISKE STUDIER I LEKSIKOGRAFI

| | |
|---|---|
| Titel: | Datamatisk leksikografi i Norden - status og visioner |
| Forfatter: | Bolette Sandjord Pedersen |
| Kilde: | Nordiske Studiar i Leksikografi 7, 2005, s. 302-314<br>Rapport frå Konferanse om leksikografi i Norden, Volda 20.-24. maj 2003 |
| 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 Sandford Pedersen*

# Datamatisk leksikografi i Norden – status og visioner

The SPINN network on harmonization of computational lexica in the Nordic countries is coming to an end, and this paper presents some of the results achieved in the project. We give a status on the computational lexica in the Nordic countries which have been represented in the network, and we look deeper into the synergies that are emerging between traditional lexicography on the one hand and computational lexicography on the other. A topic of specific interest to computational lexicography concerns the development of formal methods for the treatment of the *elasticity of word meanings*, an elasticity which in traditional lexicography is often represented via definitions and examples.

## 1 Indledning

Den datamatiske leksikografi har efterhånden eksisteret så længe – også i Nordisk sammenhæng – at det er relevant at gøre en slags status over hvor langt vi er nået, og i hvilken retning udviklingen går. Dels må vi se på status for den datamatiske leksikografis praktiske formål: de sprogteknologiske ordbaser; dels er det interessant at se på den synergi der er opstået mellem den datamatiske leksikografi på den ene side og den traditionelle leksikografi på den anden side.

NorFA-netværket SPINN (SProgteknologi og INformationssøgning i Norden) som har til formål at arbejde hen imod en harmonisering af de sprogteknologiske ordbaser for de nordiske sprog, nærmer sig sin afslutning, og artiklen omhandler bl.a. resultaterne fra dette arbejde. Interessant er det imidlertid også at se nærmere på den datamatiske leksikografis teoretiske landvindinger og på den omtalte synergieffekt mellem den datamatiske og den traditionelle leksikografi: hvad har vi lært af hinandens metoder, og hvordan kan vi yderligere udvikle dem?

Den datamatiske leksikografi tager sit udspring i datalingvistikken og er derfor traditionelt forankret i den formelle lingvistik. I praktiske sammenhænge anvendes den sprogteknologiske ordbog altid i tæt forbindelse med en sprogteknologisk grammatik – altså en formelt funderet grammatik der beskriver i hvilke sammenhænge ord indgår. Endvidere er udviklingen i det seneste tiår gået fra en interesse for generelle grammatiske regler til en hypotese om at langt mere end hidtil antaget kan uddrages af leksikalsk viden og beskrives i ordbogen – bl.a. ved hjælp af leksikalske regler. Derfor ser vi at sprogteknologiske ordbaser i dag er særdeles strukturerede og indeholder detaljeret grammatisk og semantisk viden, og vi ser også flere tiltag til at håndtere ords variabilitet i en formel sammenhæng. Dette illustreres i artiklen på basis af dels Levin og Rappaport Hovavs hypoteser om hvordan verbers syntaktiske og semantiske konstruktionspotentiale kan beskrives systematisk, dels ud fra Pustejovskys hypotese om 'det generative leksikon' som det udmøntes i de såkaldte SIMPLE-ordbøger.

De seneste landvindinger inden for korpuslingvistikken har også betydet at leksikogra-

fien generelt er blevet langt mere empirisk funderet. Både den datamatiske og den traditionelle leksikografi har nydt godt af den radikale udvikling der er sket på området, og man kan derfor sige at korpuslingvistikken har været med til at bringe de to grene af leksikografien sammen idet de nu kan tage udgangspunkt i de samme korpusanalyser. Dette eksemplificeres i artiklen ud fra Kilgarriff & Rundells arbejde med 'Word sketches'; de har udviklet og testet et korpusværktøj der gør det muligt automatisk at producere såkaldte leksikalske 'profiler'. Disse profiler produceres på baggrund af meget store korpora ved hjælp af forskellige sprogteknologiske værktøjer og gør det muligt for leksikografen på et statistisk grundlag direkte at aflæse hvilke grammatiske relationer et ord omgiver sig med.

## 2 Hvad er datamatisk leksikografi, og hvad adskiller den fra traditionel leksikografi?

Der findes mindst to fortolkninger af hvad datamatisk leksikografi rent faktisk omfatter (jf. Ooi 1999). Den ene fortolkning anser datamatisk leksikografi for at omhandle forskellige metoder til delvis automatisering af ordbogstilblivelsen. Dette kan ske ved hjælp af gode korpusværktøjer eller ved hjælp af redigeringsværktøjer som understøtter udformningen af ordbogsartiklerne bl.a. ved at konsistenschecke overbegreber eller andre elementer i ordbogsartiklen.

Den anden fortolkning, som er den vi fremover anvender i denne artikel, er at disciplinen primært omhandler opbygningen af *ordbaser til datamatiske formål*, så som avanceret tekstbehandling, maskinoversættelse, informationssøgning mv. Under opbygningen af ordbaser til datamatiske formål anvendes naturligvis i så vid udstrækning som muligt også semiautomatiske metoder og man tager så vidt muligt udgangspunkt i allerede eksisterende ordbøger (jf. Boguraev & Briscoe 1989).

Det praktiske mål for den datamatiske leksikografi i den sidste fortolkning er altså produktion af ordbaser som typisk er karakteriseret ved at bestå af strukturerede og formaliserede oplysningstyper. Dette skyldes at deres primære 'bruger' er computeren og ikke mennesket – som dog kommer ind i den næste fase, nemlig som bruger af det sprogteknologiske værktøj, fx et stavekontrolprogram. Udgangspunktet for en ordbase er – som ved udformningen af traditionelle ordbøger – typisk andre ordbøger samt korpora, men derudover er den datamatiske leksikografi som nævnt karakteriseret ved at den har et tæt tilhørsforhold til de datalingvistiske teorier der anvendes i forbindelse med formel grammatik.

Man kan altså sige at den datamatiske leksikografi skal levere ordbasen til den formelle grammatik og dette er med til at gøre at den sprogteknologiske ordbase generelt er meget styret af hvad man kunne kalde en *'top down'*-tilgang. Man anskuer altså typisk vokabularet ud fra et sæt af generelle grammatiske regler, og beskrivelser der falder uden for disse regler giver prompte problemer idet analysatoren (parseren) der anvender ordbasen i samspil med grammatikken, fejlanalyserer eller overgenererer.

Der er både fordele og ulemper ved dette afhængighedsforhold. Leksikografen tvinges til at være meget systematisk i sin ordbeskrivelse og til at følge et lukket regelsæt hvilket kan være en fordel idet ordbasen generelt bliver relativt ensartet og konsistent også selv om der er flere leksikografer indblandet. Redaktionsreglerne er med andre ord meget omfattende og præcise og består typisk af en endelig liste af mulige beskrivelser.

Ulempen ved en meget regelbaseret ordbase er imidlertid at det kan være svært at tage højde for idiosynkratiske detaljer. Fx angives der i mange traditionelle ordbøger en række brugseksempler ved især komplekse ordbeskrivelser; brugseksempler som læseren kan generalisere ud fra og derved forstå også udvidede eller specialiserede ordbetydninger. Disse udelades ofte i den sprogteknologiske ordbase fordi de er vanskelige at beskrive i et formelt sprog som computeren kan håndtere. Ordbasens prædefinerede formelle struktur kan således opfattes som en spændetrøje for leksikografen.

Som det også blev nævnt i indledningen, er udviklingen i det seneste tiår gået fra især at interessere sig for generelle grammatiske regler til en hypotese om at langt mere end hidtil antaget kan uddrages af leksikalsk viden og beskrives i ordbogen – bl.a. ved hjælp af leksikalske regler. I Chomskys tidligere arbejder var ordbogen 'a wastebin of irregularities', hos Bloomfield 'a list of irregularities, an appendix to the grammar' og holdningen var at regelmæssige variationer ikke hørte til i ordbogen som kun skulle indeholde idiosynkratiske oplysninger (Chomsky & Halle 1968). I modsætning hertil taler vi nu om *leksikalisme* udtrykt i nyere grammatiske teorier som Lexical functional grammar (LFG) hos fx Bresnan 2001 og Head-Driven Phrase Structure Grammar hos Pollard & Sag 1994. Mere og mere strukturel information indsættes i ordbasen (se figur 1) og ordet anses nu typisk for det centrale i sproget i modsætning til sætningen.

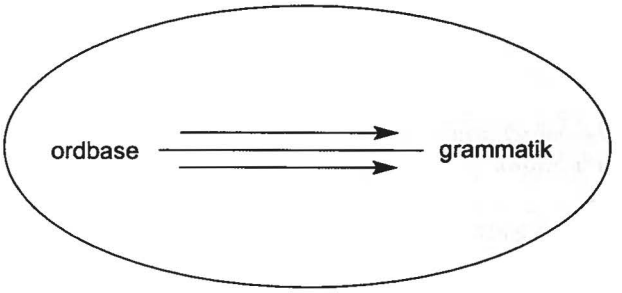
Tendensen med at anføre mere strukturel information på ordniveau ses også i traditionel leksikografi. Der er ingen tvivl om at de meget strukturerede beskrivelsesmetoder fra den datamatiske leksikografi har haft en afsmittende effekt på den traditionelle leksikografi, og at resultatet i de fleste tilfælde er et mere gennemarbejdet og systematisk slutprodukt. Det er fx efterhånden helt almindeligt at valenspotentiale angives på en mere eller mindre eksplicit måde enten ved hjælp af stiliserede eksempler eller ud fra mere eller mindre gennemskuelige forkortelser fra en lukket liste. Dette ses fx i figur 2 fra *Collins Cobuild* hvor margenoplysninger ifølge Sinclair angiver 'a link between the broad generalities of grammar and the individualities of particular words'.

Den datamatiske leksikografis svaghed mht. beskrivelse af ords idiosynkratiske og kontekstafhængige egenskaber er også et område der arbejdes på at udbedre. Flere forskellige metoder er under udvikling, dels viden/regelbaserede, dels statistiske/korpusbaserede, til at repræsentere ords variabilitet i en formel kontekst. I det følgende skal vi se nærmere på nogle af disse metoder.

## 3 Regelbaserede tilgange til beskrivelse af ords variabilitet

Levin (1993) og Rappaport Hovav og Levin (1998) og (2002) undersøger verbers variabilitet på engelsk og i Rappaport Hovav og Levin (1998) forsøger de at opstille formelle kompositionsregler for hvorfor nogle verber er meget elastiske mht. til konstruktionspotentiale, mens andre ikke er. Nogle af de spørgsmål der opstår i forbindelse med beskrivelse af verber i en ordbog, er i denne sammenhæng: hvordan beskrives verbers konstruktionspotentiale uden at skulle opregne uendelige mængder af syntaksmønstre? Hvorfor er nogle verber elastiske mens andre ikke er, og kan man generalisere over dette fænomen?

Problematikken gælder i høj grad også for de nordiske sprog, hvor nogle verber har et stort konstruktionspotentiale som det ses for dansk i eksempel 1 og 2.




Figur 1: Udviklingen går i retningen af at lægge mere information i ordbasen og mindre i grammatikken

**bury** /bɛri'/, **buries, burying, buried.** 1 When you **bury** someone who is dead, you put their body into a grave and cover it, usually with earth. eg *She will be buried here in the church... ...buried corpses.*
2 When you say that you **have buried** a particular relative, you mean that that relative has died. eg *I won't have it! I have buried enough children!... He buried his wife last week.*
3 When a person or animal **buries** something, they put it into a hole in the ground and cover it up. eg *Reptiles bury their eggs in holes or under stones... ...buried treasure.* ◆ to **bury the hatchet**: see **hatchet.**
4 When you **bury** something in a substance or under a large quantity of things, you put it there, often in order to hide it. eg *Then they buried the meat in salt... She buried the gun under a pile of leaves.*
5 If someone or something is **buried** under something that falls on top of them, for example rocks or a building, they are completely covered and often cannot get out or be reached. eg *People who had been indoors were now buried beneath mountains of rubble... Many people remained buried alive.*
6 If something is **buried** somewhere, it is beneath or behind other things where it cannot be seen and is difficult to find. eg *She found some coffee buried in the depths of her store cupboard.*
7 When you **bury** your face or head in something soft, especially in another part of your body or someone else's body, you press it against that thing so that it is hidden or partly hidden. eg *She buried her face in her hands... She stood for a moment with her head buried against his neck... She turned further away, burying her face in the pillow.*
8 When something **buries** itself somewhere or when someone **buries** it there, it is pushed very deeply in there. eg *The bullet had buried itself in the tree... You could bury your two hands in the bran.*
9 If you **bury** a particular feeling or memory of something, you try not to have or show that feeling or try to forget that thing. eg *The anger which had been buried inside me rose to the surface... ...buried memories... They agreed to bury their differences.*
10 If you **bury** yourself in a particular place, away from other people or important events, you go and spend some time there, usually alone. eg *He would voluntarily bury himself in these desert regions... ...a child who is continuously burying herself in a corner with a book.*
11 If you **bury** yourself or your head or face in

Figur 2: Opslagsordet *bury* fra *Collins Cobuild* med stiliserede valensoplysninger i margen

1)
*Peter fejede*
*Peter fejede gulvet*
*Peter fejede krummerne ind i hjørnet*
*Peter fejede bladene af fortovet*
*Peter fejede gulvet rent*
*Peter fejede bladene op i en bunke*

2)
*Peter løb*
*Peter løb til stranden*
*Peter løb en tur*
*Peter løb 2 km*
*Peter løb gaderne tynde*
*Peter løb ud/op/ned.. under/over/på.. taget*
*Peter løb sig træt*

Andre verber derimod viser sig at være meget begrænsede som illustreret i eksempel 3 og 4.

3)
*Peter smadrede en tallerken*
**Peter smadrede*
**Peter smadrede tallerknen ned af bordet*
**Peter smadrede tallerknerne op i en bunke*
**Peter smadrede gulvet fuld af tallerkner*

4)
*Peter ankom til stationen*
*Peter ankom*
**Peter ankom 2 km*
**Peter ankom en tur*
**Peter ankom sig træt*

Rappaport Hovav og Levins forklaringsmodel indarbejder verbers aspektuelle forhold og opstiller leksikalsk-semantiske skabeloner som går ud over valenspotentialet og ser på verbers semantiske klasser. Der viser sig at være en markant forskel på såkaldte *mådesverber* (manner verbs) som er meget elastiske i modsætning til *resultatverber* (result verbs) som er meget lidt elastiske. Mådesverber omfatter semantiske underklasser som bevægelsesverber (*løbe, hoppe, spadsere*) og instrumentalisverber (*feje, hamre, male*), mens resultatverber omfatter verber der iboende udtrykker en tilstandsændring så som *smadre* og *knække* samt placeringsverber som *putte* og *stille*. Rappaport Hovav og Levin opstiller leksikalske skabeloner for de 2 verbalklasser som det ses i figur 3.

**Mådesverber**
[x ACT*(MANNER)* ] ex: *løbe, hoppe, gå, spise, drikke, læse*
[x ACT*(INSTRUMENT)* ] ex: *feje, male, piske, save*

**Resultatverber**
[x CAUSE [BECOME [y*(STATE)*]]] ex: *knække, smadre, dræbe, forøge*
[x CAUSE [BECOME [y*(PLACE)* ]]] ex: *putte, stille, lægge*

Figur 3: Rappaport Hovav og Levins skabeloner for mådes- og resultatverber (1998)

[[ x ACT*(INSTRUMENT)* ] CAUSE [BECOME[y*(STATE)*]]]
ex: *Peter fejede gulvet rent*

Figur 4: En simpel event udvides til en kompleks event

For mådesverber gælder der med andre ord at en agens (x) kan handle (ACT) på en bestemt måde *(MANNER)* som tilfældet er ved bevægelsesverber og aktivitetsverber, eller ved hjælp af et bestemt instrument *(INSTRUMENT)* som tilfældet er ved instrumentalisverber. Elementerne i kursiv angiver plads til den idiosynkratiske information som fx ved *feje* er karakteriseret ved at *'rengøre el. samle noget med en kost'*. De elementer der ikke står i kursiv, angiver derimod den strukturelle information som gælder for hele klasser af verber.

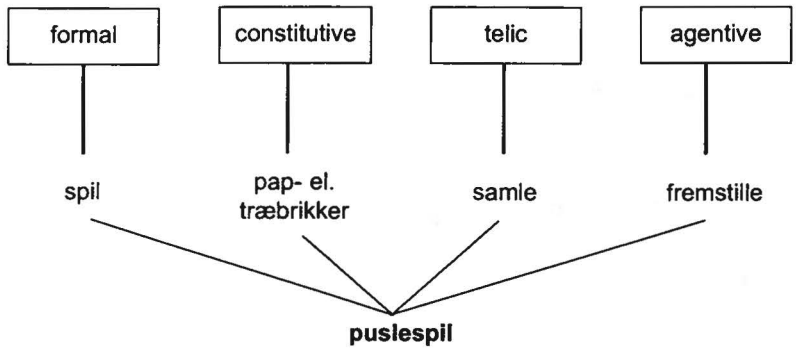
Skabelonen til resultatverber er transitiv: her afstedkommer (CAUSE) en agens (x) at noget (y) kommer i en bestemt tilstand (*STATE*) eller havner et bestemt sted (*PLACE*). Igen angiver kursiv den idiosynkratiske information som kommer fra det enkelte verbum; ved *dræbe* er tilstanden for objektet *død*; ved *forøge* er tilstanden for objektet *større end tidligere* osv.

Forskellen i elasticitet i de to klasser forklares af Rappaport Hovav og Levin med at mådesverber beskriver simple 'events' (hvilket illustreres med en simpel, ikke-indlejret skabelon) som kan *udbygges* til en kompleks event, som i figur 4.

Resultatverber er derimod allerede komplekse i og med at de har en iboende sluttilstand *(STATE)* eller *(PLACE)*, og de kan derfor ikke udbygges yderligere.

Rappaport Hovav og Levins forklaringsmodel skulle i teorien gøre det muligt i ordbogen/ordbasen blot at referere til en semantisk klasse (på samme måde som man ofte refererer til en morfologisk klasse som så udfoldes et andet sted i ordbogen) og dermed undgå at opremse en lang række valensmønstre som det ellers ville være nødvendigt hvis man skal give en udtømmende beskrivelse af bevægelsesverbers valensmønstre[1].

Pustejovsky interesserer sig i *Det generative leksikon* (Pustejovsky 1995) også for ords variabilitet. Han undersøger især hvordan substantivers semantik interagerer med de omgivende ords semantik. Han beskriver substantivers kernesemantik ud fra det han definerer som et ords *qualiastruktur*. Hvis vi ser på et ord som *puslespil* kan vi se hvordan dette ords betydning kan udtrykkes ved hjælp af de fire såkaldte *qualiaroller* som indgår i qualiastrukturen. I en traditionel ordbog som *Nudansk Ordbog* finder man følgende betydningsdefinition på ordet: *'et spil med træ- eller papbrikker i forskellige faconer som lægges sammen så de danner et hele'*. Disse betydningskomponenter kan udformes i Pustejovskys firedimensionale struktur som det ses i figur 5.




Figur 5: Qualiastrukturen for *puslespil*

Der er altså her fire betydningsdimensioner involveret for *puslespil*: (i) den formelle rolle (formal role), som giver information om ordets placering i en ontologi ved hjælp af en *is_a*-relation (svarende til genus-delen af definitionen): et puslespil er et slags spil; (ii) den konstitutive rolle (constitutive role), som udtrykker en bred vifte af semantiske relationer der typisk angår ordets interne struktur (i dette tilfælde *has_as_parts*: at det består af nogle brikker), (iii) den teliske rolle (telic role), som beskriver genstandens typiske funktion (her en *used_for*-relation: et puslepil skal samles[2]), og endelig (iv) den agentive rolle (agentive role), som beskriver genstandens oprindelse, og som primært går på om genstanden er natur- eller menneskeskabt (i dette tilfælde en *made_by*-relation).

Et ords qualiastruktur er ifølge Pustejovsky bestemmende for hvordan de såkaldte generative mekanismer arbejder når ord farver hinanden i kontekst. Ved et eksempel som *et let puslespil* kunne man fx antage at der var tale om et puslespil som ikke vejer ret meget. Den mest umiddelbare fortolkning er imidlertid at der er tale om et puslespil som er let at lægge, dvs. at *let* her betyder *'som ikke volder nogen større problemer eller anstrengelser'* og den mekanisme der træder i kraft, er det som Pustejovsky kalder *selektiv binding* (selective binding). Ved at inferere ud fra puslespils teliske rolle, kan der foretages en korrekt entydiggørelse af *let*[3] nemlig betydningen *let at samle.* Også i mange andre tilfælde af flertydighed er der en tendens til at formålet med kulturskabte ting spiller en meget vigtig rolle når vi fortolker de ord der omgiver substantivet, og formålet med kulturskabte entiteter tildeles derfor en særlig vægt i de generative mekanismer[4].

En anden generativ mekanisme, *typeændring* (type coercion), træder ifølge Pustejovsky bl.a. i kraft ved overførte betydninger. *Puslespil* optræder ofte i overført betydning som i eksempel 5 og 6:

5) *at opklare forbrydelsen var et puslespil*
6) *de forskellige løsfund danner tilsammen et puslespil over 100 års rigmandsliv*

I disse tilfælde ændrer *puslespil* type fra at være en konkret kulturskabt enhed (artifakt) til at betyde *'en kompleks sag som består af flere dele som skal samles til et hele'* altså et begreb

| | Konkret | Overført | I ordbøger | Formål |
|---|---|---|---|---|
| *vindue* | 92 % | 8 % (15) | nej | *se* |
| *våben* | 90 % | 10 % (100) | nej | *kæmpe* |
| *bro* | 75 % | 25 % (75) | ja | *forbinde* |
| *bombe* | 50 % | 50 % (150) | nej | *ødelægge* |
| *panser* | 40 % | 60 % (10) | ja | *beskytte* |
| *nøgle* | 30 % | 70 % (274) | ja | *åbne* |
| *piedestal* | 25 % | 75 % (12) | ja | *placere højt* |
| *spændetrøje* | 20 % | 80 % (34) | ja | *fastholde* |
| *puslespil* | 20 % | 80 % (67) | nej | *samle* |
| *glidebane* | 20 % | 80 % (12) | nej | *glide* |
| *rygstød* | 11 % | 89 % (16) | ja | *læne* |
| *vifte* | 10 % | 90 % (72) | nej | *afkøle* |
| *narresut* | 8 % | 92 % (11) | ja | *trøste* |
| *sovepude* | 0 % | 100 % (14) | ja | *sove på* |
| *skyklapper* | 0 % | 100 % (14) | ja | *afskærmning* |
| *springbræt* | 0 % | 100 % (38) | ja | *sætte af* |

Figur 6: Den procentvise hyppighed af konkrete og overførte betydninger i korpus ved en række artifakter på dansk (Nimb & Pedersen 2000:682)

af typen abstrakt. Det er imidlertid interessant at betydningskomponenten *'samle til et hele'* synes at gå igen i den overførte betydning og faktisk definere denne. Også ved overførte betydninger spiller den teliske rolle altså en afgørende rolle.

I Nimb & Pedersen (2000) undersøges denne tendens mere systematisk i korpus med en række danske artifakter der også anvendes hyppigt i overført betydning, som det ses i figur 6.

Figur 6 illustrerer flere interessante forhold. For det første er det markant hvor mange relativt hyppige overførte betydninger der ikke er opført i traditionelle mellemstore ordbøger, fx *vindue*, *våben, bombe* og *puslespil* for ikke at nævne *vifte* hvoraf kun 10% af forekomsterne er konkrete. Derudover viser tendensen med den teliske rolles betydning for ords variabilitet sig at være meget tydelig. For næsten samtlige overførte betydninger i figur 6 er den teliske rolle fra den konkrete betydning central for tolkningen af den overførte betydning. *En bombe* er noget der potentielt kan *ødelægge*, *våben* bruges til at *kæmpe* med, *et puslespil* skal *samles* etc. Ganske vist tolkes den teliske rolle også overført; det er altså *samle* i overført betydning vi taler om når detaljerne omkring en forbrydelse skal lægges sammen så de danner et hele osv. En interessant undtagelse fra reglen er *vifte* hvor det snarere er den konstitutive rolle der refereres til i den overførte betydning; her refererer vi til måden en vifte er sat

sammen på af en række fjer, træ- eller papstykker som foldes ud. De enkelte dele repræsenterer i den overførte betydninger muligheder eller alternativer som det ses i eksempel 7 og 8:

7) *Vi har en meget bred vifte af tilbud*
8) *Der er blevet igangsat en bred vifte af aktiviteter*

Med udarbejdelsen af *Det generative leksikon* forestiller Pustejovsky sig ideelt set at den i princippet endeløse opremsning af under- og overbetydninger bliver unødvendig i den datamatiske ordbase. Ords qualiastruktur i samspil med sprogets generative mekanismer er ideelt set tilstrækkeligt som beskrivelsesgrundlag for ords variabilitet. Leksikografer der arbejder konkret med udviklingen af sprogteknologiske ordbaser, har nok erfaret at sagen er mere kompliceret end som så og at ords idiosynkratiske egenskaber har større kraft end der er lagt op til i Pustejovskys teori.

Alligevel er teorien interessant som forklaringsmodel og den danner da også grundlag for det store, europæiske ordbaseprojekt SIMPLE (Semantic Information for Multifunctional PLurilingual LExica) hvor man bl.a. ønsker at nå frem til standarder for semantisk ordbeskrivelse også i et multilingvalt perspektiv (jf. Lenci et al. 2000 for en beskrivelse af projektet som helhed, og Pedersen & Paggio (2004) samt Pedersen & Keson 1999 for den danske SIMPLE-ordbase). I SIMPLE-projektet indgår qualiastrukturbeskrivelsen i ordbogen og alle betydningsbeskrivelser refererer til en fælles top-ontologi.

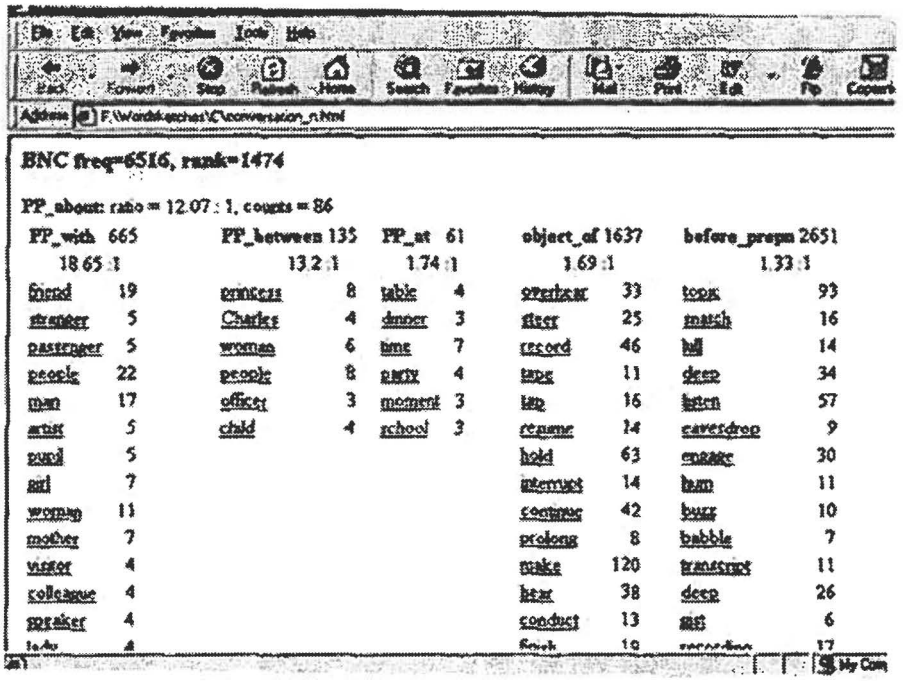
## 4 Korpusbaserede tilgange til beskrivelse af ords variabilitet

Vi har før nævnt eksemplets funktion i traditionelle ordbøger som en kilde til læseren til selv kreativt at komplettere ordets betydningspotentiale. I den sammenhæng spiller korpusleksikografien oplagt en central rolle. En af de nyere landvindinger på dette områder angår anvendelsen af såkaldte *leksikalske profiler* (word sketches) til ud fra store korpora at identificere ords variabilitet og kombinatoriske potentiale.

I dag er det fast rutine ved de fleste ordbogsprojekter at leksikografen manuelt gennemgår og eventuelt opmærker en sorteret korpuskonkordans for at danne sig et indtryk af et ords konstruktions- og betydningspotentiale. Evt. anvender man mere raffinerede værktøjer som kan undersøge ords indbyrdes tiltrækningskraft (mutual information statistics) således at man kan se hvilke ord der hyppigt optræder sammen set i relation til hvor hyppige ordene i det hele taget er i korpus (Church and Hanks 1989). Muligvis har man et ordklasseopmærket korpus således at man kan sortere støj fra ved absolutte homografer og søge mere præcist på specifikke kombinationer af ord eller ordklasser. Alligevel kræver denne form for korpusanalyse en hel del introspektion og en vis frasortering af støj, altså irrelevante eksempler.

Ved udarbejdelsen af leksikalske profiler (Kilgarrif & Rundell 2002) har man gennemarbejdet korpusset yderligere ved hjælp af nogle sprogteknologiske værktøjer. Udover at være ordklasseopmærket er korpusset også parset, dvs. man har ladet det undergå en grovkornet automatisk syntaksanalyse. Efter denne proces udgør korpusset en langt mere pålidelig kilde for statistiske beregninger på grammatiske relationer så som subjekt, objekt og præpositionsobjekt. Beregning på grammatiske relationer giver yderligere den fordel at flertydighed bliver lettere at diagnosticere.

I figur 7 ses den leksikalske profil for det engelske substantiv *conversation*.




Figur 7: Leksikalsk profil for conversation (fra Kilgarrif & Rundell 2002:812)

Den første og anden kolonne angiver de ord der udfylder styrelsen i de valensbundne præpositionsforbindelser indledt med hhv. *with* og *between*, som i *a conversation with my friend* og *conversations between Princess Diana and Prince Charles*. Kolonne 3 angiver typisk steds- og tidsadverbialer indledt med *at* som i *a conversation at the table* eller *a conversation at this moment*. Ordene er ordnet efter relativ frekvens, dvs. *people* optræder længere nede i første kolonne end *friend* simpelthen fordi *people* i hele korpus er mere frekvent end *friend*. I kolonne 4 anføres de verber som *conversation* typisk er objekt for; altså *overhear a conversation* og så fremdeles. Ordene i kolonnerne fungerer som links til de korpuseksempler de kommer fra, så leksikografen har mulighed for løbende at afklare og verificere konteksten.

Ved hjælp af leksikalske profiler vil leksikografen i fremtiden have langt bedre mulighed for at kortlægge de generelle mønstre i et sprogs vokabular. Hvis en gruppe ord har samme profil, har vi fx et godt grundlag for at antage at de også ligner hinanden rent semantisk. Dette kan vi måske i fremtiden udnytte til en ny og bedre strukturering af vores leksikalske data. Leksikalske profiler kan i fremtiden også komme til at betyde at den traditionelle og den datamatiske leksikografi kommer til at nærme sig hinanden yderligere simpelthen fordi man måske vil benytte samme velfunderede udgangspunkt for sin beskrivelse.

## 5 Status for SPINN-netværket

SPINN-netværket (SProgteknologi og INformationssøgning i Norden) nærmer sig sin afslutning. Netværket har eksisteret i nu 3 år (2001-2003) og har haft stor tilslutning fra Island, Norge,

Danmark og Sverige til sine aktiviteter i den forløbne periode (jf. http://www.cst.dk/spinn/spinn-home.html).

Det primære formål med netværket har været at tage udfordringen op mht. harmonisering og standardisering af ordbaser på de nordiske sprog for således at bane vejen for igangsættelse af ordbogsafhængige sprogteknologiske applikationer for disse. Især har der været fokus på ordbaser til brug for mere indholdsbaseret søgning for de nordiske sprog (se Pedersen et al. 2001).

Det før omtalte EU-projekt SIMPLE dannede udgangspunkt for SPINN-netværket, idet både de danske og svenske SPINN-koordinatorer havde deltaget i projektet og opbygget ordbaser for hhv. dansk og svensk. Dette projekt har også dannet grundlaget for netværkets arbejde omkring sammenhægtning af ordbaser for de nordiske sprog via den såkaldte SkanLex-model (se Pedersen et al. 2003: kap. 4), således at fx tværsproglig søgning mellem de nordiske sprog muliggøres. SPINN- netværket blev endvidere igangsættelsen af norsk SIMPLE.

Status for de tre SIMPLE-ordbøger på hhv. norsk, svensk og dansk er som følger:

- Norsk SIMPLE: norske ækvivalenter til dansk SIMPLE er udarbejdet (Ruth Fjeld m.fl.), der arbejdes nu også med svenske ækvivalenter til norsk og dansk samt med databasestruktur og med kodningsværktøj (Preben Wik m.fl.);
- Svensk SIMPLE: Man arbejder med en semi-automatisk udvidelse af ordbasen (Maria Toporowska, Dimitrios Kokkinakis m.fl.).
- Dansk SIMPLE: fortsættes i STO-regi, altså i sprogteknologisk ordbase for dansk (se Braasch & Pedersen 2002); dog med en noget 'lettere' semantik i første omgang.

Også andre ordbaseressourcer (udover SIMPLE-ressourcer) har været involveret i SPINN, bl.a. NordKompLex (Torbjørn Nordgaard m.fl.), Svensk Ordnet (Åke Viberg m.fl.), Ordnet for norsk (Helge Dyvik m.fl.) og terminologibaser (Bodil Nistrup Madsen m.fl.). Herudover har der været samarbejde med NorFA-netværket om navnegenkendelse, Nomen Nescio (Janne Bondi Johannessen m.fl.).

Netværkets udadvendte aktiviteter har bestået i 1-2 fællesmøder årligt for de i alt ca. 30 netværksdeltagere; nogle med inviterede internationale gæster (bl.a. Nicoletta Calzolari, Carol Peters, Jussi Karlgren) samt tre ph.d.-kurser i almen og datamatisk leksikografi i hhv. 2002, 2003 og 2004 med undervisning af bl.a. Michael Rundell, Adam Kilgarriff, Helge Dyvik, Gregor Thurmair og Allessandro Lenci.

Den seneste udvikling i netværket er at vi er blevet inviteret med i det europæiske netværk ENABLER (European National Activity for Basic Language Resources (www.enabler-network.org)) hvis formål er dels at udvide samarbejdet mellem forskellige nationale projekter som omhandler udviklingen af sprogteknologiske ressourcer, dels at identificere fælles sprogteknologiske mål og dermed bedre udnytte synergieffekter. Yderligere formål med netværket er at værne om sprogressourcernes kompatibilitet for at muliggøre overførsel af sprogteknologier og værktøjer mellem forskellige sprog samt at stimulere udviklingen af flersproglige ressourcer. Netværket arbejder desuden for at kortlægge tilgængelige sprogressourcer og skabe et internationalt overblik over disse (International Roadmap for Language Resources).

## 6 Konkluderende bemærkninger

Det er vigtigt at datamatisk leksikografi i Norden holder trit med den internationale udvikling på området, og det er vigtigt for de nordiske sprog at der udvikles tidssvarende sprog-

teknologiske værktøjer ikke bare for engelsk, men også for vores modersmål. En forudsætning for sådanne værktøjer er gode sprogressourcer bl.a. i form af ordbaser.

I denne sammenhæng ser vi på længere sigt flere fordele ved at SPINN indgår i ENABLER-netværket nu hvor NorFA-bevillingen slutter. Dels giver dette samarbejde mulighed for at synliggøre det nordiske arbejde inden for sprogteknologi på det europæiske landkort og samarbejdet med andre landes forskere fremmes; dels kan samarbejdet lette adgangen til eksisterende sprogressourcer på europæisk plan. Det er også særdeles relevant for udviklingen i Norden at vi er med til at etablere *standarder* på området og dette kræver internationalt samarbejde.

I fremtiden lader det til at *The Semantic Web* vil få stor betydning for vores adfærd på internettet. Semantic Web betegner en ny form for web-indhold som er forståelig for maskiner, og som skal gøre det muligt at søge mere indholdsorienteret på internettet bl.a. ved hjælp af semantiske metadata, jf. Berners-Lee et al. 2001. Det synes at være en forudsætning for en positiv udvikling af konceptet i The Semantic Web at der skabes synergi med sprogteknologien, og dette indebærer bl.a. at man indarbejder sprogteknologiske ordbaser i den nye teknologi (jf. også Lenci et al 2003).

## Referencer


Berners-Lee, T., J. Hendler & O. Lassila 2001. «The Semantic Web». *Scientific American,* maj 2001 http://scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2.

Boguraev, B. & E.J. Briscoe (eds.) 1989. *Computational Lexicography for Natural Language Processing*, London and New York: Longman.

Braasch, A. & B. Pedersen 2002. «Recent Work in the Danish Computational Lexicon Project STO», in *EURALEX Proceedings 2002,* Center for Sprogteknologi, Copenhagen.

Bresnan, J. 2001. *Lexical Functional Syntax*, Blackwell Textbooks in Linguistics, Blackwell Publishers, Mass. USA.

Chomsky, N. & M. Halle 1968. (apud Ooi. 1998). *The sound pattern of English,* New York: Harper and Row.

Church, K. & P. Hanks 1989. «Word association norms, mutual information and lexicography», in: *ACL Proceedings,* Vancouver:76-83.

*Collins Cobuild English Language Dictionary.* 1987. Collins Publishers & University of Birmingham.

Kilgarriff, A. & M. Rundell 2002. «Lexical Profiling Software and its Lexicographical Applications – a Case Study», in: *The Tenth EURALEX 2002 Proceedings, pp. 807-818,* København.

Lenci, A. N. Calzolari 2003. «From Text to Content: Computational Lexicons and the Semantic Web». Uudgivet artikel uddelt ved SPINN-seminar (kan downloades fra http://cst.dk/spinn/oslophdpraes.html)

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, in: T. Fontenelle (ed.) *International Journal of Lexicography Vol 13.* 249-263. Oxford University Press.

Levin, B. 1993. *English Verb Classes and Alternations, A Preliminary Investigation.* The University of Chicago Press, Chicago and London.

Nimb, S. & B.S.Pedersen 2000. «Treating Metaphoric Senses in a Danish Computational Lexicon – Different Cases of Regular Polysemy» in: *The Ninth EURALEX 2000 Proceedings, pp. 679-691,* Stuttgart.

Ooi, V.B.Y. 1998. *Computer Corpus Lexicography*, Edinburgh University Press.

Pedersen, B., R.V. Fjeld, M. Toporowska Gronostaj 2001. «Harmonisering og sammenkædning af sprogteknologiske ordbaser med særligt henblik på informationssøgning – en rapport fra SPINN-netværket», i: *NorFAs Årsskrift 2001.*

Pedersen, B., R.V. Fjeld, M. Toporowska Gronostaj 2003. «Sprogteknologiske ordbaser for de nordiske sprog – rapport fra et forskningsnetværk». *Skrifter udgivet af Nordisk Forening for Leksikografi Nr. 7:273-291*, Torshavn, Færøerne.

Pedersen, B., Paggio, P. 2004. «The Danish SIMPLE Lexicon and its Application in Content-based Querying», in *Nordic Journal of Linguistics Vol. 27 no. 1.* 97-127.

Pedersen; B.. & Britt Keson 1999. «SIMPLE – Semantic Information for Multifunctional Plurilingual Lexica: Some Danish Examples on Concrete Nouns», in: *SIGLEX99: Standardizing Lexical Resources, Association of Computational Linguistics, ACL99 Workshop pp.*, Maryland.

*Politikens Nudansk Ordbog* 1999. Politikens Forlag A/S.

Pollard, C. & I. Sag 1994. *Head-Driven Phrase-Structure Grammar.* The University of Chicago Press, Chicago & London.

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

Rappaport Hovav, M. and B. Levin 2001. «An Event Structure Account of English Resultatives», *Language* 77, 766-797.

Rappaport Hovav, M. & B. Levin 1998.«Building Verb Meanings». In,: M. Butt W. Geuder (eds.) *The Projection of Arguments: Lexical and Compositional Factors*, CSLI Publications, Stanford, CA.

## Noter

1. Specielt i den sprogteknologiske ordbase synes dette at give problemer idet man traditionelt angiver fx alle verbets mulige partikler og efterfølgende mulige præpositioner. Kombinationspotentialet synes nærmest uendeligt; især ved bevægelsesverber.
2. Bemærk at den mere overordnede funktion af et puslespil måske snarere er at underholde eller at opøve visse logiske eller geometriske færdigheder end blot det *at samle.* Disse mere overordnede formål med spil nedarves imidlertid automatisk fra ordet *spil* som *puslespil* forbindes til via den formelle rolle.
3. I *Nudansk Ordbog* har *let* følgende betydninger: *(i) som ikke vejer el. fylder ret meget* (fx taske: fysisk genstand) *(ii). som har en lav styrke el. grad* (fx trafik: fænomen) *(iii). som ikke volder nogen større problemer el. anstrengelser, el. som sker uden at man gør noget særligt* (fx om tilberedning: handling) *(iv) som er ubekymret el. overfladisk* (fx om person). Definitionen på *puslespil* er i *Nudansk Ordbog* som følger: *et spil med træ- el. papbrikker i forskellige faconer som skal lægges sammen så de danner et hele.* Andre dimensioner i kernesemantikken for *puslespil* er således at det har *spil* som overbegreb og består af flere *pap-* eller *træbrikker.*
4. Pustejovsky bruger selv eksemplerne *a fast car* og *a fast typist* til at illustrere selektiv binding. *Fast* går i begge tilfælde på den teliske rolle, altså hhv. *køre* og *taste.*