LexicoNordica

17 · 2010

LEKSIKOGRAFI OG
SPRÅKTEKNOLOGI
I NORDEN

NORDISK FORENING FOR LEKSIKOGRAFI

LexicoNordica

LexicoNordica

17 · 2010

LEKSIKOGRAFI OG
SPRÅKTEKNOLOGI
I NORDEN

NORDISK FORENING FOR LEKSIKOGRAFI

LexicoNordica 17 · 2010
Leksikografi og språkteknologi i Norden

Omslag og sats: Laurids Kristian Fahl
Trykk: Rosendahls – Schultz Grafisk A/S

LexicoNordica trykkes med støtte fra
Ekspertgruppen Nordens språkråd

ISSN 0805-2735

INNHOLD

Leksikografi og språkteknologi i Norden

Ruth Vatvedt Fjeld & Henrik Lorentzen

Årets tema i LexicoNordica er "Leksikografi og språkteknologi i Norden". På det forberedende symposiet på Schæffergården ble det holdt 11 faglige foredrag som ble avsluttet med en sammenfattende diskusjon. Fem av de nordiske landene var representert. På symposiet ble det særlig drøftet hvilke deler av våre fagområder som er av felles interesse for språkteknologi og leksikografi, og i hvilken grad de to fagfeltene kan ha nytte av hverandres forskning og kartlegging. Disse foredragene er nå utarbeidet til faglige artikler og utgjør årets nummer av LexicoNordica sammen med to ikke-tematiske bidrag og åtte meldinger av forskjellige nordiske leksikografiske utgivelser. I tillegg er en artikkel en kommentar til en tidligere artikkel i LexicoNordica, og til sist meldinger om de nordiske leksikografikonferansene i 2009 og 2011.

Lenge har man i leksikografisk arbeid nå hatt stor praktisk nytte av datamaskinell lagring og bearbeiding av sine store datamengder, og det er fortsatt svært viktig. Ikke minst er gode redigeringsprogrammer en uvurderlig hjelp. Men slike programmer er ofte forretningshemmeligheter, og de har vært lite eller svært overflatisk omtalt. Det er jungeltelegrafen som har vært den viktigste informasjonskanalen når et leksikografisk prosjekt skulle velge redigeringsprogam, og det er betydningsfulle valg for leksikografiske prosjekter, som ofte innebærer et stort og varig arbeid med store kostnader. Artikkelen til Sjögreen & Sköldberg gir en god oversikt over ordboksredigeringsprogrammer som er i bruk i Norden. Artiklene til Halskov og Bick presenterer også egentlig hjelpeprogrammer for tradisjonell leksikografi, men går ut over det vi vanligvis tenker på som redigeringsprogram.

Men disse artiklene er sånn sett litt på siden av årets hovedtema, som representerer noe relativt nytt innen leksikografien, nemlig hva slags leksikalsk beskrivelse man har behov for når man lager maskinleselige leksikon.

Forholdet mellom språkteknologi og leksikografi er svært tett i moderne tid, det er nesten utenkelig å bedrive leksikografi i dag uten å nytte språkteknologiske ressurser. På den andre siden er språkteknologien blitt avhengig av gode leksikalske beskrivelser, langt på vei er disse to fagdisiplinene gått inn i en symbiose. Vi ser det som nyttig at fagfolk fra begge leire møtes og diskuterer muligheter og samarbeid framover, også mellom de nordiske landene.

Flere av artiklene gir et oversyn over status quo for språkteknologi og leksikografi pr. januar 2010, dvs. hvilke språkteknologiske programmer og ressurser for leksikografisk beskrivelse som er tilgjengelige. Det gjelder Pedersen fra Danmark, Borin og Kann fra Sverige, Rögnvaldsson fra Island, og Hagen & Nøklestad fra Norge. Slike oversikter er gode utgangspunkt for videre arbeid, for mulig samarbeid og for arv av andres metoder og produktutvikling, slik at man unngår dobbeltarbeid. De nordiske språkene har så mye til felles at slikt arvegods er svært verdifullt. Andre artikler presenterer spesielle språkteknologiske verktøy som er til stor hjelp i leksikografisk beskrivelse, som Bicks kollokasjonsanalyseprogram, andre språkteknologiske prosjekter der den leksikografiske beskrivelsen er særs viktig, som Lindén & Carlson og Pedersen om WordNet, Halskov om automatisk nyordsregistrering og Nikulasdóttir & Whelpton om automatisert semantisk beskrivelse av substantivers betydning. I artikkelen til Trosterud legges det særlig vekt på hvor viktig det er at språkteknologer og leksikografer samarbeider, særlig i små språksamfunn der kostnadene ved leksikalsk dokumentasjon relativt sett blir store.

Til sammen gir disse artiklene en god oversikt over hvor langt forskningen i grensefeltet mellom språkteknologi og leksikografi er kommet i Norden i dag, og vi får en del konkrete beskrivelser

av hvilken praktisk nytte man kan ha av det i dokumentasjon av ordforråd, og også viktigheten av leksikografisk beskrivelse i andre språkteknologiske produkter.

Det er to ikke-tematiske artikler i denne utgaven av LexicoNordica. En artikkel dokumenterer arbeidsmetoden for store leksikografiske verk, der redaktøren sitter med en mengde ordbokssedler og redigerer uten hjelp av moderne, språkteknologiske analyser, nemlig Wendts beskrivelse av den tradisjonelle redigeringsmåten i store, dokumenterende ordbøker. Den andre ikke-tematiske artikkelen, skrevet av Pálfi, Stokholm & Tarp, gir en historisk oversikt over ordbøker mellom dansk og ungarsk.

Det er en viktig oppgave for et skrift som LexicoNordica at det også presenterer ordbøker og andre leksikografiske produkter som kommer ut, og gir kritiske anmeldelser av dem. Det har vi absolutt klart i årets årsskrift, med 10 artikler som handler om utgitte ordbøker eller andre leksikografiske prosjekter. Slike meldinger er nødvendige kritiske refleksjoner rundt midler og metoder i leksikografien, som driver faget videre og øker bevisstheten om faget.

LexicoNordica er også et forum for faglig diskusjon og meningsbryting, noe Becker-Christensens kommentar til et tidligere bidrag er et godt eksempel på.

Som vanlig inneholder også denne utgaven av årsskriftet meldinger fra Nordisk forening for leksikografi.

V

LexicoNordica har med dette nummer fått ny redaksjon, med to nye hovedredaktører og også utbytting av noen andre medlemmer i redaksjonen. Den nye redaksjonen består av Ruth Vatvedt Fjeld og Henrik Lorentzen som hovedredaktører, og Sturla Berg-Olsen, Ken Farø, Jón Hilmar Jónsson, Nina Martola og Emma Sköldberg som nasjonale redaktører. I alle tidligere nummer helt fra 1994 har Henning Bergenholtz og Sven-Göran Malmgren vært hovedre-

daktører. De har hatt en god hånd med årsskriftet og lagt opp en løype som gjør det rimelig greit å fortsette. Etter et års arbeid som hovedredaktører har vi imidlertid også fått enda større respekt for det store arbeidet som de i så mange år har utført for det nordiske leksikografiske fagmiljøet og fellesskapet.

Det store arbeidet med layout har tidligere vært ivaretatt af Kristinn Jóhannesson. Det er fra og med dette nummer overtatt av Laurids Kristian Fahl. Tidsskriftet har i denne forbindelse også fått en ansiktsløftning med nye skrifttyper og en mer luftig sats. Vi håper det faller i lesernes smak.

Symposiet på Schæffergården i januar mottok økonomisk støtte av Stiftelsen Clara Lachmanns fond, av Letterstedtska föreningen og av Fondet for dansk-norsk samarbejde. Utgivelsen av dette nummer av årsskriftet har fått tilskudd fra Ekspertgruppen Nordens språkråd. Redaksjonen takker hjertelig for denne støtten.

Neste års symposium blir som vanlig holdt på Schæffergården utenfor København 28.-30. januar 2011. Videre informasjon om symposiet legges ut på foreningens hjemmeside (http://www.nordisk-sprakrad.no/nfl). Temaet for de to kommende symposiene er fastlagt av hele redaksjonen. De blir som følger:

2011 Onomasiologiske ordbøker i Norden
2012 Betydningsbeskrivelser i nordiske ordbøker

Til symposiet i 2012 mottar vi gjerne forslag til foredragsholdere, som kan sendes til vedkommende lands redaktør. Redaksjonen mottar ellers også gjerne forslag til temaer for kommende symposier.

Til sist vil vi takke lederen i Nordisk forening for leksikografi Birgit Eaker for alt arbeid hun har lagt ned i å søke om midler som gjør det mulig å holde symposier og utgi årsskriftet. Og aller sist vårt sine qua non, Rikke Hauge fra Språkrådet i Norge, som holder

i alle tømmer og ender slik at både det praktiske og sosiale rundt symposier og redigeringsarbeidet går så glatt som mulig.

Ruth Vatvedt Fjeld
professor
Institutt for lingvistiske og nordiske studier
Universitetet i Oslo
Postboks 1001 Blindern
NO-0315 Oslo
r.e.v.fjeld@iln.uio.no

Henrik Lorentzen
seniorredaktør, cand.mag.
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København K
hl@dsl.dk

TEMATISKE BIDRAG

DeepDict – et korpusbaseret relationelt leksikon

Eckhard Bick

DeepDict (at www.gramtrans.com) is a new type of lexical resource, built from grammatically analysed corpus data. Co-occurrence strength between mother-daughter dependency pairs is used to automatically produce dictionary entries of typical complementation patterns and collocations, in the fashion of an instant monolingual usage dictionary. DeepDict is capable of abstracting lemma relations and semantic classes from inflected surface forms, and provides concordances and statistics for the relations found. Entries are supplied to the user in a graphical interface with various thresholds for lexical frequencies as well as absolute and relative co-occurrence frequencies. DeepDict draws its data from Constraint Grammar-analysed corpora, ranging between tens and hundreds of millions of words, covering the major Germanic and Romance languages, among them both Swedish, Danish and Norwegian. Apart from its obvious lexicographical purposes, DeepDict also targets teaching environments and translators.

1. Leksikografisk motivation

I bred leksikografisk forstand vil en korpusbaseret ordbog ikke alene generelt have et bedre dækningspotentiale, men også en større autenticitet end en traditionel ordbog kompileret vha. introspektion og litterære citater. Mange moderne ordbøger gør derfor brug af korpusdata, optimalt set med udgangspunkt i et materiale, der er balanceret mht. domæne, register etc. Alligevel ligner slutproduktet, den publicerede ordbog, som regel stadigvæk en traditionel ordbog, selv i elektroniske udgaver, fordi korpusdata er blevet brugt mere til eksemplificering, eller i bedste fald

frekvensoplysninger, end til egentlige ordbogsopslag. To undtagelser er *Sketch Engine* (Kilgariff et al. 2004), der benytter sig af n-gram-kollokationer og grammatiske relationer på systematisk vis, og *Wortschatz*-projektet ved Universität Leipzig (Biemann et al. 2004), der genererer netværk af semantisk beslægtede ord fra monolingvale korpora.

Men selv hvor der benyttes korpora i det leksikografiske arbejde, det være sig selektivt eller systematisk, kan der være store begrænsninger i tilgængeligheden af den information, der gemmer sig i et korpus, især hvad angår strukturel information, fordi de fleste korpora af den nødvendige størrelse kun foreligger som rene tekstkorpora, uden dybere grammatisk opmærkning. Allerede det mest basale opmærkningsniveau, med lemmatisering og ordklasse-entydiggørelse, vil tillade en bedre udnyttelse af korpusmaterialet, normalisering og optælling svarende til opslagsordets grundform etc.; men først en dyb syntaktisk-funktionel opmærkning med markering af subjekts- og objektsrelationer m.m. tillader ekstraktion af strukturelle relationer mellem ord, der ikke står umiddelbart ved siden af hinanden i teksten (såkaldte n-grammer).

Endelig, selv hvor leksikografen har adgang til et opmærket korpus af tilstrækkelig størrelse, med en brugerflade, der tillader opstilling af konkordanser og ordstatistik, vil det kun være muligt at undersøge ét relationelt mønster ad gangen – en besværlig proces, ikke mindst for verber med et komplekst frasalt og semantisk konstruktionspotentiale. Og ofte kan et givent mønster slet ikke findes i korpusset, enten fordi søgeformalismen ikke er tilstrækkelig finkornet, idet den fx er tekstbaseret snarere end kategoribaseret, eller fordi korpora med den nødvendige opmærkningsdybde (en såkaldt træbank) som regel kun produceres som håndopmærkede korpora med få hundredetusinde ord[1].

1 Karel Kalurand anfører netop begrænsninger af denne type, dvs. dækningsgrad og statistisk prægnans, som problemer i forbindelse med hans

Det leksikografiske redskab, der præsenteres her, DeepDict, forsøger at gå nye veje, både hvad angår den lingvistiske kvalitet i den tilgængelige korpusinformation, og mht. en mere integreret præsentation af de relationelle informationer for det enkelte ord. DeepDict blev udviklet af GrammarSoft Aps og lanceret på internetadressen www.gramtrans.com i september 2007.

I modsætning til en papirordbog har en elektronisk ordbog som DeepDict ingen volumenbegrænsninger, så opslaget for et sjældent ord kan fylde lige så meget som for et højfrekvent ord, og udelukkelsen af sjældne ord og relationer behøver derfor ikke at være absolut, men kan reguleres af brugerstyrede tærskler. Men særlig store bliver fordelene for en produktionsordbog: På papirmediet er det nemlig nemmere at fremstille passive ("definitions-") ordbøger end aktive (produktivt-kontekstuelle) ordbøger, fordi førstnævnte henvender sig til modersmålsbrugere af målsproget (MS), mens sidstnævnte optimalt set skal levere en stor mængde detaljerede brugsinformationer, semantiske restriktioner og kompletteringsmønstre for brugere med MS som fremmedsprog. Fx "A gives x to B" – med A, B som person-variable (+HUM) og x, y som ting-variable (-HUM). En elektronisk ordbog kan derimod rumme et væld af brugsinformation "on demand" og tilbyde ubegrænsede korpuseksempler – eksempler, der ikke optager plads i det primære opslag og først bliver synlige, når brugeren aktiverer et tilsvarende link.

2. Kompileringen af en leksiko-relationel database

For at honorere de krav om robust og detaljeret grammatisk korpusopmærkning, der blev drøftet i kapitel 1, valgte vi Constraint Grammar (CG, Karlsson et al. 1995) som sprogligt analyse- og

deepdict-lister, der bygger på en estisk CG-baseret træbank med 100.000 ord (http://math.ut.ee/~kareel/NLP/Programs/Treebank/DepDict).

opmærkningsparadigme, dels pga. metodens meget lave parsing-fejlprocenter og gode leksikalsk-morfologiske dækningsgrad, dels fordi CG-syntaksen bygger på dependensrelationer, dvs. relationer mellem ord snarere end mellem non-terminale konstituenter, med al syntaktisk information tilgængelig på ordniveau – et forhold, der medfører betydelige lettelser i computer-processeringen af opmærkede data. I det følgende beskrives den valgte fremgangsmåde for opbygningen af en leksiko-relationel database.

2.1. Korpusopmærkning

Det første skridt for hvert sprog bestod i den grammatiske opmærkning af samtlige tilgængelige korpora vha. Constraint Grammar-parsere, efterfulgt af en dependens-analyse med CG-tags (fx @SUBJ = subjekt, @ACC = direkte objekt) som input (Bick 2005). Resultatet kan beskrives som en gigantisk træbank på ca. en milliard ord, med dependensrelationer for samtlige ord i hver sætning[2]. For nogle af vores korpora var det dog kun det sidste trin, der var del af DeepDict-projektet selv, idet materialet allerede forelå som CG-opmærkede korpora inden for CorpusEye-systemet (http://corp.hum.sdu.dk). Tabel 1 giver et overblik over art og omfang af de anvendte korpora.

I det nedenstående opmærkede sætningseksempel har både subjektet *Peter* (ord 1) og objektet *nødder* (ord 6) dependensrelationer (#x→y) til verbet *spiste* (ord 2).

Peter "**Peter**" <hum> PROP @**SUBJ** #1→2
spiste "spise" V IMPF #2→0
en håndfuld ....
nødder "**nød**" <fruit> N P @**ACC** #5→2

2 Dependenstræerne har fuld dybde og er således informationsækvivalente med tilsvarende konstituent-træbanker, CG3-dependenser (beta.visl.sdu.dk/constraint_grammar.html) eller Functional Dependency Grammar (www.connexor.fi).

	Korpus-størrelse[3]	Genre	Parser[4]	Status[5]
Dansk	159 mio.	blandet	DanGram	+
Engelsk	210 mio.	blandet	EngGram	+
Esperanto	58 mio.	blandet	EspGram	+
Fransk	[67 mio.]	Wiki, Europarl	DTT+FrAG	–
Italiensk	46 mio.	Wiki, Europarl	DTT+ItaGram	+
Tysk	44 mio.	Wiki, Europarl	GerGram	+
Norsk	50 mio.	Wiki, kundedata	Obt / NorGram	+
Portugisisk	210 mio.	avis, Europarl	PALAVRAS	+
Spansk	90 mio.	internet, wiki, Europarl	HISPAL	+
Svensk	60 mio.	avis, Europarl	SweGram	+

Tabel 1: Korpora og parsere

2.2. Dependensbigrammer

Det er denne type binære relationer, dvs. dependenspar, der blev "høstet" fra de opmærkede korpora, med informationer om lemma, ordklasse og syntaktisk funktion for både dependenten ("datterordet") og hovedet ("moderordet").

Peter_SUBJ → spise_V
kat_SUBJ → spise_V
nød_ACC → spise_V
mus_ACC → spise_V

For at undgå en eksplosion af informationsløs leksikalsk mangfoldighed blev talord og navne udelukkende gemt uden deres lemma, for sidstnævnte dog med en markering af semantisk klasse,

3 Wiki = Wikipedia (http://www.wikipedia.com), Europarl = the Europarl Corpus (Koehn 2005).

4 Mere information om parserne fås på: http://beta.visl.sdu.dk/constraint_grammar.html.

5 Der er fri adgang til DeepDict for portugisisk, svensk og esperanto, mens der kræves login/abonnement for de øvrige sprog.

fx <hum> (menneske), <org> (organisation) etc. Også præpositioner fik en særbehandling i ekstraktionsprocessen; dels var det styrelsen, dvs. den semantiske kerne, snarere end præpositionen selv, der blev betragtet som hovedet, dels blev der de facto brugt 3-leds-relationer, idet præpositioner blev gemt som en slags kasusmarkør sammen med deres styrelse (fx *tygge* ← *på* ← *problem* giver relationen *tygge* ← *problem**på*).

De fleste af de anvendte parsere leverer foruden den syntaktiske også en semantisk opmærkning med såkaldte semantiske prototyper for substantiverne – i stil med den allerede nævnte navneklassificering, men på et højere distinktionsniveau med ca. 200 prototyper. <fruit> (frugt), for eksempel, er en undertype af <food> (mad), der igen kan være en undertype (<food-c>, <food-m>) af <cc> (tællelige konkreta) eller <cm> (mængdekonkreta). En række hovedkategorier tilføjer semantiske underklasser som små bogstaver efter et stort bogstav for hovedklassen, fx <Vair> (air vehicle), <tool-cut> (skære-redskab) og <Hprof> (human professional).

Lægger man de enkelte lemma-, ordklasse- og prototype-relationer samlet ind under dependenshovedet som opslagsord, får man fx for verbet *eat* ('spise') et summarisk opslag, der viser, hvem der spiser (SUBJ-subjekt, fx PROP-proprium), og hvad der spises (ACC-objekt):

{PROP, *kat*, <hum>, ...} SUBJ → *spise*
spise ← {*nød*, *mus*, <fruit>} ACC

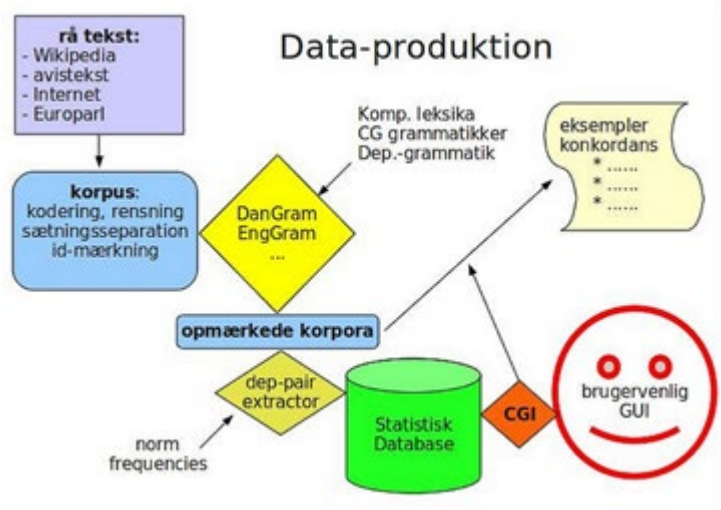
2.3. En database over korrelationsstyrker

Det er åbenlyst, at dependentlisterne i et sådant opslag uden statistisk information hurtigt ville blive reduceret til "leksikalsk støj" af den kombinatoriske mangfoldighed i et stort korpus. Det er med andre ord nødvendigt at skelne mellem typiske komplementer og korrelationer på den ene side og ikke-informativ "støj"-variation

på den anden side. Vi har derfor benyttet et statistisk mål for korrelationsstyrke, dvs. sandsynligheden for samforekomsten af 2 ord i en given syntaktisk relation. For at sondre mellem typiske og ikke-informative korrelationer dividerede vi den absolutte frekvens for samforekomsten med produktet af korpus-normalfrekvenserne for hvert af de 2 ord alene:

$$C * \log(p(a \rightarrow b) \, ^2 / \, (p(a) * p(b)))$$

hvor p() står for frekvenser, og C er en konstant, der sammen med logaritmiseringen blev introduceret for at placere statistisk signifikante værdier mellem 0 og 10. Forskellen mellem vores formel og Church's *Mutual Information*-mål (Church & Hanks 1990) er den øgede vægtning (^2 = kvadratvægtning) af selve samforekomstfrekvensen – en vægtning, vi anså for gavnlig i leksikografisk øjemed, fordi den hindrer stærke men sjældne eller forkerte kollokationer i at udkonkurrere kollokationer bestående af mere almindelige ord (og tilsvarende høje frekvensværdier i brøken).



Figur 1: Data-produktion og GUI (graphical user interface)

Den endelige standardiserede dependensdatabase indeholder for hvert "dep-gram"-ordpar, foruden dets absolutte frekvens og kookkurens-styrke, også et indeks over id'erne på de relevante sætningsforekomster i kildekorpusset.

Selv for et enkelt sprog kan hele processen tage dage eller uger, og databaserne har en størrelse (p.t. 90 GB), der gør det umuligt at benytte sig af almindelige database-programmer, fordi et enkelt opslag ville medføre en for brugeren uacceptabel ventetid på flere minutter, og vores interface-programmør, Tino Didriksen, var nødt til at udvikle særlige opslagsalgoritmer og multiple filstrukturer for at løse problemet.

3. Brugerinterfacet

Opslag i DeepDict er dynamiske "leksikogrammer" – frekvenssorterede, grafisk ordnede lister af kollokater. Præcis hvilke kollokater der vises, er afhængigt dels af opslagsordets ordklasse og dermed typiske funktionelle kompletteringsmønster, dels af en række tærskelværdier, der kan sættes individuelt for at tilgodese forskellige brugerprofiler:

- minimum-forekomst (af dependens-kollokationen) – bruges til at bortfiltrere tekstfejl, opmærkningsfejl og hapaxer
- minimum-kookkurrens-styrke (default > 0) – regulerer typiciteten af kollokaterne
- maksimum antal kollokater, der vises per funktionsfelt
- leksikalsk minimumsfrekvens for kollokat-ordene (4 niveauer) – kan bruges til at sikre, at kun almindelige ord vises som kollokater, fx til skolebrug

Af grammatiske årsager skelnes mellem fx "tale_V" (verbum) og "tale_N" (substantiv/nomen), og hver ordklasse har sin egen leksikogramskabelon. Leksikogrammet for det engelske substantiv

voice, for eksempel, indeholder således ikke bare typiske flerordsudtryk som *voice actor* eller *voice recorder*, men viser også typiske attributter (i feltet "premodifier"), fx *loud*, *deep*, *husky* og det flertydige *passive voice*.

Figur 2: Substantiv-leksikogram

Felterne i DeepDict er placeret på en måde, der understøtter "naturlig læsning". Attributter findes derfor til venstre og hoveder til højre for adjektiviske og substantiviske opslagsord på engelsk, svarende til sprogets normale ordfølge. Tilsvarende placeres subjekter til venstre for et verbum og objekterne til højre. Nogle felter er forsynet med en tekstramme for at skabe illusionen af en "sætning", fx "one can *{recognize, hear, lower, lend, raise}* a voice".

Værdierne for kookkurrensstyrken angives optionelt som røde tal foran det enkelte kollokatord, efterfulgt af en kolonseparator og den duale logaritme af den absolutte forekomst. Som default vises kun kollokationer med en logaritmeklasse på 2 eller højere (4 eller flere forekomster). Rækkefølgen af ordene i et felt er en

samlet funktion af kookkurensstyrke og absolut frekvens, og for yderligere at skelne mellem sikre og usikre kollokater vises høje logaritmeklasser med fed skrift. Når man klikker på et kollokat, åbnes et konkordansvindue der viser sætningseksempler og en fuldformsstatistik for den pågældende lemma-kollokation.

Forms	Abs Freq	Rel Freq
Total	6749	100.00%
vægt -> lægger	3608	53.46%
vægt -> lægge	1270	18.82%
vægt -> lagde	820	12.15%
vægt -> lagt	815	12.08%
vægten -> lægger	100	1.48%
vægten -> lægge	68	1.01%

Show all forms...

word sketch:

lægger . stor mest afgørende meget **vægt** **på** {. / . / , fordi } {bag / på denne / på ved ansættelser / på det}

Concordances for: vægt_N -> lægge_V

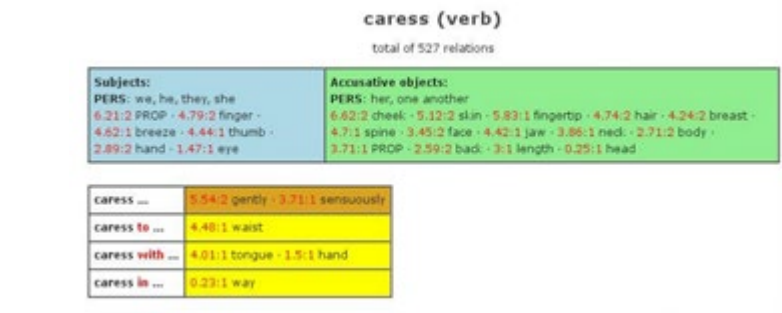
ID	Text
inf30-33978	« Markedsføring på værdier og ansvar skal være en vigtig kilde til at øget konkurrenceevne på verdensmarkedet , fordi den globale forbruger i stadig større udstrækning **lægger vægt** på den slags , når han eller hun køber ind .
inf100-35624	I sidste ende giver den diplomatiske balancegang dog mulighed for , at den tyrkiske regering for en tid kan **lægge** mest **vægt** på sin egen tillægserklæring af indenrigspolitiske årsager , mens EU fokuserer på toldaftalen .
inf40-148330	« I dag **lægger** vi megen **vægt** på , at tilldsmanden skal arbejde med uddannelse og kunne svare på spørgsmål om pension .
c2000-	Netop den pind har vi fået ind , og det **lægger** vi stor **vægt** på , fordi vi netop vil gøre

Figur 3: Konkordansopslag

For støtteverbumskonstruktioner kan det være nødvendigt med en dependensdybde større end 2, dvs. at vise flere komplementer på én gang, som i udtrykket *lægge ... vægt på/bag ngt*. Her fungerer ordet *vægt* som syntaktisk objekt, men indgår i en inkorporation med verbet, hvis egentlige komplement er præpositionssyntagmet *på ...*. Mens DeepDicts primære opslag kun fokuserer på det umiddelbare objekt, vises hele konstruktionen i konkordansopslaget som en såkaldt "word sketch".

Personlige og kvantitative pronominer er så frekvente, at eksakte statistiske værdier her kun har begrænset interesse. Til gengæld kan pronominer levere semantisk information, "abstraheret"

som pronominale prototyper (fx ±human, køn, ±tællelig, sted/retning), og DeepDict viser derfor en ordnet liste af karakteristiske pronominer på subjekts- og objektspladserne. Personlige subjektspronominer kan hjælpe med at klassificere aktiviteter som typisk mandlig ('han') eller kvindelig ('hun'), markere objekter som mængdeord ('meget') eller endda tillade sociolingvistiske deduktioner. Således viser DeepDict-opslaget for det engelske verbum *caress* at mænd ('he') typisk er subjekt og kvinder ('her') typisk objekt i kærtegningsrelationen.



Figur 4: Verbums-leksikogram

Eksemplet viser desuden, at metaforisk brug dækkes ind på samme måde som konkret brug – således vises der ud over objekt-kropsdele, der kærtegnes, og subjekt-kropsdele, der kærtegner (*finger*, *thumb*), også metaforiske agenter som *breeze* og *eye*. Endelig illustreres, hvordan præpositioner (*with tongue/hand*) håndteres i DeepDicts verbalskabelon.

Adverbium-verbums-kollokationer eksisterer i flere funktionelle varianter – (a) ubundne tids-, steds- og mådesadverbier, (b) valensbundne adverbier (feel *how*, go *where*) og (c) verbalintegrerede partikler (give *up*, fall *apart*), og i nogle tilfælde kan det endda være svære at skelne mellem kategorierne (fx cut *out*). Fordi formålet med DeepDict er leksikografisk snarere end syntaktisk, nøjedes vi dog her med kun at fremhæve verbalpartiklerne som

separat klasse (for at understøtte en underlemmatisering af det pågældende verbum) og at samle alt andet adverbielt materiale i en og samme paraplykategori (brunt felt, fx *gently/sensuously* for verbet *caress*).

4. Betydningsnuancer igennem dependens-kollokater

Selvom DeepDict også for polyseme ord viser kollokaterne samlet[6], kan det undertiden hjælpe at udgrænse forskellige kernebetydninger, nemlig igennem det semantiske spektrum af kollokaterne (fx både konkrete og abstrakte prototyper) og igennem den syntaktiske funktion, der knyttes til en given relation. Således fremgår det af leksikogrammet for det portugisiske adjektiv *pesado* ('tung'), at ordet både anvendes konkret (= 'af høj vægt') og abstrakt (= 'betydelig/alvorlig'), og at det i førstnævnte betydning har en tendens til at blive brugt som postmodifikator, mens det foranstilles som præmodifikator ved abstrakte kollokater.

Figur 5: Adjektiv-leksikogram

6 Medmindre distinktionen allerede er en del af den forudgående korpusopmærkning.

Tilsvarende kan DeepDict hjælpe med at fremhæve betydningsnuancerne mellem nære synonymer. Således kan det for en studerende af dansk som fremmedsprog være svært at vide, hvornår han skal benytte hhv. *mistænksom* og *mistænkelig*. De to tilsvarende opslag på DeepDict vil imidlertid gøre det klart, at førstnævnte beskriver et udtryk/indtryk, mens sidstnævnte bruges om handlinger og hændelser:

mistænksom ...	**mistænkelig ...**
stemme, ?forsvindingsnummer, receptionist, øje, grimasse, blik, gemyt, tonefald, ?transaktion	transaktion, person, færden, grad, pengeoverførsel, adfærd, personage, dødsfald, forhold
<expression>	<action, event, situation>

Tabel 2: Betydningsnuancer

Omvendt kan det for en dansker være vanskeligt at anvende de engelske adjektiver *big*, *large* og *high* korrekt, men også her formår DeepDict-korrelaterne implicit at "definere" betydningerne:

high ...	**big ...**	**large ...**
• level • [school] • concentration • speed • proportion • altitude • elevation • temperature	• [bang, band] • hit • problematic • break • difference • brother • star, bird • man, city	• number • quantity • amount • proportion • sum • portion, part • city, island • population
<degree> <measure>	<size> <importance>	<extension> <quantity-mass>

Tabel 3: Semantisk motiverede kollokationsrestriktioner

Samtidigt identificeres visse flerordsudtryk *[big bang, big band]*, engelske komposita med tryk på første led. Men mens sådanne

flerordsudtryk også er tilgængelige for en ren tekstuel kollokationsanalyse, drager de øvrige, funktionelle kollokationer fordel af CG-dependensrelationerne. Således vil relationen *high + temperature* findes, selv hvis der ikke foreligger en eneste sætning, hvor ordene står ved siden af hinanden – fordi relationerne også fanges i *high room temperature* eller i prædikativ brug, *ambient temperature was rather high when ….*

I et bilingvalt perspektiv kan DeepDict advare brugeren om, at en oversættelse, selv mellem nært beslægtede sprog, ikke nødvendigvis matcher ordene en-til-en. Således rummer den svenske oversættelse *smeka* af dansk *kærtegne* også betydninger (fx 'stryge'), der end ikke metaforisk dækkes af det danske ord, og DeepDict-leksikogrammet viser dette igennem de fundne typiske objekter:

kærtegne ...	**smeka ...**
• bryst, krop, kind, hud, balder, mave, inderlår, brystvorte, hår, ansigt, klitoris, lår, sexbombe, nosse, røvhul, nakke, hals, kropsdel, bagdel • silkestof, græsbane • PROP-hum	• kind, könsorgan, bröst, stjärt, klitoris, kropp • PROP-hum • boll, passning, tennisboll • elgitarr • lack, rännil, instrumentpanel, julle, murbrok, vidunder

Tabel 4: Bilingval polysemikontrol

5. DeepDict som arbejdsredskab

Eksemplerne i de forudgående kapitler viser art og omfang af den information, der gemmer sig i DeepDict-opslagene. Men på sin vis er der tale om et uslebent værktøj, hvor mange muligheder nok understøttes, men på den anden side forudsætter en vis grad af nytænkning og tilpasning hos brugeren. Oplagte brugergrupper ud over den almindelige "ordbogs"-bruger er (a) leksikografen

og (b) universitetsunderviseren. Leksikografen kan således finde inspiration mht. kompletteringsmønstre, flerordsudtryk, frasale verber m.m. og uddrage de mest karakteristiske eksempler for en given konstruktion, snarere end bare de mest frekvente. Bl.a. vil en metaforisk kombination ofte udvise en høj korrelationsværdi, netop hvis den ene part ellers er et lavfrekvent ord. Desuden understøttes semantiske subdistinktioner og sammenligninger som vist ved adjektiveksemplerne i sidste afsnit.

For underviseren kan DeepDict, i forbindelse med udarbejdelse af det relevante didaktiske materiale, være et middel til at stimulere de studerendes sproglige nysgerrighed og give undervisningen et mindre teoretisk, men mere empirisk og datanært præg, især når redskabet kombineres med almindelig korpusbrug. Mulighederne strækker sig fra ordfelt-øvelser (fx mad & drikke, via verberne *spise* og *drikke*, sprog- og landenavne etc.), over kombinatoriske undersøgelser (hvilken præposition styres typisk af et givent substantiv eller verbum?) til semantiske (fx metaforer) eller sociolingvistiske øvelser (fx konnotationerne af ordene *udlænding*, *indvandrer* og *flygtning* igennem tilknyttede adjektiver).

6. Konklusion og perspektivering

DeepDict viser, hvordan syntaktisk relaterede ordpar kan "høstes" fra grammatisk opmærkede dependenskorpora til at kompilere en statistisk database, der tillader genereringen af såkaldte "leksikogrammer" – halvgrafiske oversigtssider for monolingvale ordbogsopslag, med information vedrørende hoved- og modifikator-selektionsrestriktioner, verbalkomplettering og frasale kollokationer. DeepDict gør det muligt for leksikografen ikke alene at finde korpuseksempler og -frekvenser for bestemte (kendte) kollokationer og leksikale strukturer, men også at kompilere (nye) lister over sådanne kollokationer og strukturer.

6.1. Bedre parsere

Rent forskningsmæssigt kan de statistiske informationer fra DeepDict-databasen bruges til at forbedre CG-parserne, der så igen kan levere bedre korpora til en ny runde DeepDict-generering. Således har forfatteren udvidet det portugisiske parsingleksikon med tags for sandsynligheden for at en given syntaktisk funktions-"slot" udfyldes af en bestemt semantisk prototype:

- *pensar* ('tænke'): <fSUBJ/H:74>, <FSUBJ/org:25>
  (<fSUBJ/H:74>: f=frekvens, SUBJ=subjekt, H=human, 74=frekvensprocent)
- *competir* ('konkurrere'): <fPRP-com/H:81>, <fPRP-com/A:18>

Denne type information kan så bruges i fx en anaforgrammatik til at human-markere portugisiske personlige pronominer, der ellers kun har grammatisk køn:

ADD (£hum) TARGET PERS + @P<
(p @PIV LINK 0 PRP-COM LINK p (<fPRP-com/H>70>))

(markér PERS som human[£hum], hvis den fungerer som styrelse (@P<) til et præpositionelt objekt (@PIV) 'com' (=med), som så igen har et dependenshoved (p) med verbal-kompletterings-tag der kræver både præpositionen 'com' og trækket H (human) med en sandsynlighed større end 70)

6.2. Framenet

DeepDicts nuværende leksikogrammer fokuserer på én binær relation ad gangen, dvs. at fx subjektsfeltet og objektsfeltet beregnes uafhængigt af hinanden. Mens dette er fuldt tilstrækkeligt til mange anvendelser, kan det i en fuldstændig beskrivelse af verbets potentiale være interessant også at inddrage mulige gensidige

afhængigheder af subjekter og objekter og derfor at arbejde med såkaldte "frames" (http://framenet.icsi.berkeley.edu/), fx <hum> 'læse' <sem-r>, i stedet for dependens-bigrammer (<hum> 'læse' og 'læse' <sem-r> hver for sig). Dette kan imidlertid lade sig gøre med de samme annoterede korpora som udgangspunkt, og forfatterens plan er således at benytte DeepDicts database til at fuldføre det påbegyndte danske framenet på www.framenet.dk.

6.3. Brugertilpasning

Med slutbrugere i tankerne kan DeepDict som integreret eller separat modul kobles til andre leksikale ressourcer – traditionelle definitionsordbøger, ontologier eller bilingvale ordbøger (fx QuickDict-ordbøgerne på www.gramtrans.com), hvor DeepDict kan udfylde rollen som *aktiv* ordbog, dvs. vise brug og brugsrestriktioner for et givet målsprogsord.

Fordi DeepDict-metoden i princippet er anvendelig for alle typer af tekstkorpora, der kan analyseres med en Constraint Grammar-parser, vil det desuden være muligt at forsyne sprogforskere, leksikografer og lærere med individuelle DeepDict-installationer for specifikke brugerkorpora, tilpasset et bestemt domæne, en særlig genre eller forskellige geografiske eller sociale sprogvarianter.

Litteratur

Bick, Eckhard 2005: Turning Constraint Grammar Data into Running Dependency Treebanks. I: Civit, Montserrat & Kübler, Sandra & Martí, Ma. Antònia (red.): *Proceedings of TLT 2005*, Barcelona, December 9th–10th, 2005), 19–27.

Bick, Eckhard 2006: A Constraint Grammar-Based Parser for Spanish. I: *Proceedings of TIL 2006 – 4th Workshop on Information and HLT.*

Biemann, Chris & Stefan Bordag & Uwe Quasthoff & Christian Wolff 2004: Language-Independent Methods for Compiling Monolingual Lexical Data. I: *Comp. Linguistics and Intelligent Text Processing*. Berlin: Springer, 217–228.

Church, Ken & Patrick. Hanks 1990: Word Association Norms, Mutual Information and Lexicography. I: *Computational Linguistics,* vol.16:1, 22–29.

Karlsson, Fred et al. 1995: Constraint Grammar – A Language-Independent System for Parsing Unrestricted Text. I: *Natural Language Processing*, no. 4. Berlin & New York: Mouton de Gruyter.

Kilgarriff, Adam, Rychlý, P., Smrž, P. & Tugwell, D. 2004: The Sketch Engine. I: *Proceedings of Euralex 2004 (Lorient, France)*, 105–116.

Koehn, Philipp 2005: Europarl – A Parallel Corpus for Statistical Machine Translation. I: *MT Summit X* (Sept.12–16, 2005). Phuket, Thailand.

Eckhard Bick
forskningslektor, dr.phil.
Syddansk Universitet
Rugbjergvej 98
DK-8260 Viby J
eckhard.bick@mail.dk

Med Zipf mot framtiden – en integrerad lexikonresurs för svensk språkteknologi[1]

Lars Borin

Digital resources resulting from research projects will often languish once the project ends. Lack of funding for resource maintenance, resource non-interoperability and closed-content license formats contribute to this. At the Swedish Language Bank, University of Gothenburg, we are now integrating a number of existing free lexical resources into a new open-content resource for Swedish language technology applications. We ensure interoperability among resources by using the standardized SALDO lexicon sense and lemgram identifiers as pivot. On top of the integrated resource, we are defining a Swedish framenet, reusing a considerable amount of linguistic knowledge already encoded in the existing resources.[2]

1. Bakgrund

Språkresurser – (annoterade) korpusar, grammatiker och lexikon – är centrala i all språkteknologi. De tillhandahåller det språkliga köttet och blodet i de praktiska tillämpningarna, men de är också oundgängliga för metodutvecklingen, eftersom språkresurserna – framför allt annoterade korpusar – används som 'facit' för de metoder man strävar efter att utveckla för att skapa nya resurser, helst

1 Det arbete som beskrivs nedan har kunnat utföras tack vare finansiellt stöd av Vetenskapsrådet (i projektet *Framtidssäkring av Språkbanken* 2008–2010 – VR dnr 2007-7430) och Göteborgs universitet, dels genom ordinarie anslag till Språkbanken, dels inom *Centre for Language Technology* (CLT – <http://www.clt.gu.se>) med strategiska medel tilldelade styrkeområdet språkteknologi 2009–2012.

2 See <http://spraakbanken.gu.se/eng/saldo> and <http://spraakbanken.gu.se/eng/swefn>.

med så liten mänsklig inblandning som möjligt.

En viktig anledning till att man vill utveckla automatiska metoder för att bygga upp språkresurser är att dessa resurser kräver mycket stora arbetsinsatser för sitt förverkligande. Detta gäller i synnerhet lexikonresurser, som är den resurstyp som vi ska koncentrera oss på här.

Lexikalisk kunskap har kommit att inta en alltmer central plats i språkteknologin. Till en del beror detta på en parallell utveckling inom lingvistiken, där alltmer av den kunskap som vår språkförmåga antas bygga på har klassificerats om från grammatisk till lexikalisk, en utveckling som tog fart på 1980-talet och som förknippas med grammatikformalismer som LFG, GPSG och HPSG.[3] Ett sätt att tänka på detta är som ett skifte i fokus: Tidigare antogs syntaktiska konfigurationer vara primära. De tillhandahöll positioner där lexikonord av lämplig typ kunde hämtas ur lexikonet och stoppas in (”lexikal insättning”). I den s.k. ”radikala lexikalism” som kännetecknar LFG och dess efterföljare uppstår istället de syntaktiska konfigurationerna som ett resultat av samspelet mellan lexikonenheterna, medan själva den syntaktiska komponenten nu på sin höjd omfattar några få mycket allmänna frasstrukturregler.

Många projekt har kommit till stånd för att bygga språkteknologiska lexikonresurser för diverse språk, både från maskinläsbara lexikon för mänskligt bruk och från grunden. Idealt ska en lexikonresurs för språkteknologi innehålla all relevant lingvistisk information om ord och flerordsenheter, alltså information om dessa enheters morfologi, betydelse, pragmatik, uttal och ämnesområdestillhörighet, samt om deras syntaktiska och semantiska

3 LFG: Lexical-Functional Grammar (uppsatserna i Bresnan 1982, särskilt Bresnan & Kaplan 1982); GPSG: Generalized Phrase Structure Grammar (Gazdar et al. 1985); Autolexical syntax (Saddock 1991); HPSG: Head-driven Phrase Structure Grammar (Pollard & Sag 1994). Det fanns även tidigare ansatser i denna riktning, av vilka *Word Expert Parsing* (Small & Rieger 1982) särskilt förtjänar att nämnas.

kombinerbarhet, allt detta uttryckt på ett så formellt sätt att all denna information kan användas för automatisk bearbetning av texter. Samtidigt ska en sådan lexikonresurs vara omfattande nog att kunna användas i applikationer som arbetar med obegränsad text (eller tal), samt med lätthet kunna kopplas ihop med (likaledes formellt uttryckt) omvärldskunskap. Man kan nog lugnt konstatera att sådana lexikonresurser inte existerar idag.

Å andra sidan finns (ibland ansenliga) fragment av alla dessa typer av information, men utspridda över flera resurser som har tagits fram i olika projekt vid olika tidpunkter av olika forskargrupper. Här handlar det både om digitalisering av lexikon för mänskligt bruk och nyskapande av lexikonresurser specifikt för språkteknologianvändning.

Eftersom dessa befintliga resurser representerar stora insatser i tid och pengar och eftersom de i många fall innehåller högvärdig språklig information, har vi i Språkbanken vid Göteborgs universitet startat ett projekt för att rädda så mycket som möjligt av våra egna existerande digitala lexikonresurser från förgängelsen samt vidareutveckla dem.[4] Det förra består huvudsakligen i att integrera existerande resurser, det senare handlar främst om att till den integrerade resursen lägga den typ av semantisk och syntaktisk information om lexemen som man finner i det engelska Berkeley FrameNet (Johnson & Fillmore 2000) och några få liknande resurser för andra språk (Boas 2009), men även om att komplettera de befintliga resurserna med flerordsenheter. Det tilltänkta slutresultatet går under arbetsnamnet *Svenskt frasnät++* (eng. ”Swedish FrameNet++”: *SweFN++*), där ”++” signalerar att resursen redan från början kommer att innehålla betydligt mer information och även mer varierad information än bara frasnätet. Speciellt kan nämnas att SweFN++ planeras som en diakronisk resurs, alltså att

4 Projektgruppen består för närvarande av Lars Borin, Dana Dannélls, Markus Forsberg, Annika Kjellandsson, Dimitrios Kokkinakis och Maria Toporowska Gronostaj.

vi i den kommer att integrera lexikonresurser som beskriver flera olika historiska stadier av svenska. Se vidare avsnitt 2 nedan.

I denna uppsats ska jag främst beskriva vårt arbete med att integrera de befintliga resurserna. Arbetet med frasnätsinformationen beskrivs närmare på annan plats (Borin et al. 2010) och kommer inte att beröras i detalj här.

Följande principer är vägledande för integrationsarbetet:

Interoperabilitet: De resurser som står till vårt förfogande har kommit till vid olika tidpunkter och för olika ändamål. Först under senare år har insikten om vikten av standardisering på allvar börjat slå igenom i språkteknologiforskargemenskapen, något som avspeglas bl.a. i bildandet av en ISO-kommitté för språkresursstandardisering.[5] Integrering innebär följaktligen för oss inte bara att de befintliga resursernas format och innehåll anpassas inbördes, utan även – kanske viktigare – att resultatet blir 'framtidssäkert' så att det kan återanvändas i många olika sammanhang genom att vi använder oss av befintliga och framväxande standarder. Se vidare avsnitt 3 nedan.

Öppet innehåll: Vårt mål är att SweFN++ ska bli en fri lexikonresurs för svensk språkteknologi. Med "fri" menar vi att den görs tillgänglig under en licens som gör den till öppen källkod/öppet innehåll (Open Source/Open Content). Det för dock med sig att alla resurser som vi bakar in i SweFN++ också måste vara tillgängliga under samma typ av licensvillkor. I avsnitt 2 nedan ges en kort karakteristik av ett antal sådana fria lexikonresurser, både sådana som vi har utarbetat i Språkbanken och sådana som har tagits fram av andra.

Metodutveckling: Med begränsade ekonomiska och personella resurser är det orealistiskt att tro att vi ska kunna nå vårt mål – att SweFN++ förutom att integrera huvuddelen av de resurser som beskrivs i nästa avsnitt, även ska innehålla frasnätsinformation för

5 ISO TC 37/SC 4 (Language resource management); se <http://www.tc37sc4.org/>.

50.000 lexikonenheter – med enbart manuellt arbete. Ett uttryckligt mål i projektet är således att skapa ett arbetsflöde där automatiska metoder och befintliga språkteknologiverktyg används i största möjliga utsträckning, och manuellt arbete sätts in enbart där det är absolut oundgängligt och/eller där det ger mest utdelning för insatsen. Metodologiska aspekter av vårt arbete diskuteras i avsnitt 4 och 5 nedan.

2. Existerande fria lexikonresurser

2.1. Resurser i Språkbanken

SALDO kommer att utgöra 'navet' i SweFN++ och alla andra resurser länkas via SALDO. Resursen innehåller lexikalisk-semantisk och morfologisk information om 73.000 betydelser och är därmed den omfångsrikaste av våra fria resurser.[6] SALDO har beskrivits utförligt i andra publikationer (Lönngren 1989; Borin 2005; Borin et al. 2008; Borin & Forsberg 2009a) och läsaren hänvisas till dessa för detaljer.

De svenska PAROLE- och SIMPLE-lexikonen har utvecklats inom de två EU-samarbetena PAROLE (1996–1998) och SIMPLE (1998–2000) (Lenci et al. 2000). PAROLE-lexikonet innehåller 29.000 syntaktiska enheter med syntaktisk valensinformation. SIMPLE-lexikonets 8.500 betydelser är försedda med information om semantisk typ, ämnesområde, urvalsrestriktioner och vilken syntaktisk enhet i PAROLE-lexikonet som realiserar betydelsen. Dessa två resurser och SDB (se nästa stycke) innehåller en mängd information som kommer att vara direkt återanvändbar vid definitionen av frasnätets semantiska och syntaktiska ramar.

6 <http://spraakbanken.gu.se/saldo>

Semantisk databas (SDB) tillhandahåller valensbeskrivningar för ett antal verb med användning av en semantisk rolluppsättning innehållande ungefär 40 allmänna semantiska roller (Järborg 2001). Valensbeskrivningarna är vidare länkade till förekomster av verben i en balanserad korpus (ungefär 200.000 instanser), vilket alltså i praktiken utgör ett svenskt ordbetydelsedisambiguerat korpusmaterial.

Dalins ordbok (Dalin 1850–53) avspeglar språket vid mitten av 1800-talet och innehåller ungefär 63.000 uppslagsord. Den har digitaliserats av Språkbanken och är tillgänglig för sökning av uppslagsord via ett webbgränssnitt.[7] Hopkopplingen av Dalin och SALDO på betydelsenivå pågår inom ett separat e-vetenskapsprojekt (Borin, Forsberg & Kokkinakis 2010). I skrivande stund har knappt 47.000 uppslagsord ur Dalin länkats automatiskt till SALDO som ett första led i det arbetet.[8] Språkformen i Dalin är klart skild från det moderna språket (bl.a. genom en mellanliggande stavningsreform), men ändå så pass närstående detta att vi tror att integreringen inte kommer att bereda några större problem.

Språkbankens fornsvenska lexikonresurser består i tre digitaliserade ordböcker över fornsvenska (1225–1526): Söderwall 1884 och 1953, samt Schlyter 1887. Tillsammans innehåller de tre lexikonen ungefär 25.000 ingångar (men även en stor mängd sammansättningar och flerordsenheter listade under huvuduppslagsorden). I ett tidigare projekt har vi definierat en basmorfologi för fornsvenska som inkluderar den stavningsvariation som faktiskt iakttas i fornsvenskt textmaterial (Borin & Forsberg 2009b). I kontrast mot 1800-talsspråket i Dalins ordbok står fornsvenskan dock mycket långt från det moderna språket. Av den anledningen kommer de fornsvenska resurserna inte att integreras i SweFN++ i

7 <http://spraakbanken.gu.se/dalin/>

8 Se <http://spraakbanken.gu.se/eng/research/swefn/dalin/statistik>.

första omgången, men på längre sikt ser vi det som en spännande metodologisk och teoretisk utmaning att ta oss an detta arbete.

2.2. Fria lexikonresurser från andra källor

Folkets synonymlexikon – Synlex (Kann & Rosell 2006; se även Kanns uppsats i denna volym) – är resultatet av ett kollektivt wikipedialiknande initiativ där en stor mängd användare av nätversionen av det engelsk-svenska Lexin-lexikonet har ombetts bedöma graden av synonymi hos ett ordpar (slumpmässigt valt ur en stor mängd synonymkandidater) på en skala från 0 till 5. Den nedladdningsbara versionen av lexikonet innehåller alla ordpar med bedömningen 3 eller högre, närmare 40.000 ordpar.[9] Genom att koppla ihop Synlex, SALDO och lexikalisk-semantiska relationer ur SDB, bygger vi nu ett slags ordnät för svenska – Swesaurus – som kommer att innehålla både graderade synonymer och SALDOs associationsrelationer i en och samma resurs. Hittills har vi kopplat ungefär 8500 monosema uppslagsord i Synlex till SALDO (Borin & Forsberg 2010).[10]

Intercontinental Dictionary Series (IDS) och **Loanword Typology (LWT)** är ordlistor skapade för forskning i lexikal typologi (Koptjevskaja-Tamm et al. 2007) med ungefär 1.800 betydelser som antas ges lexikalt uttryck i ett stort antal språk.[11] Dessa fritt

9 Se <http://lexikon.nada.kth.se/synlex.html>.

10 Vissa relevanta resurser är tyvärr omöjliga att använda i detta arbete. Sedan ett antal år tillbaka existerar början till en svensk version av Princeton WordNet. Det svenska ordnätet är dock inte tillgängligt på villkor som skulle tillåta oss att införliva det i vår integrerade resurs, som vi ju planerar att göra fri under en öppen källkodslicens. Istället räknar vi tyvärr med att behöva skapa motsvarande information själva helt från början. På samma sätt har Brings svenska motsvarighet till Rogets tesaurus (Bring 1930) digitaliserats två gånger i två olika projekt, men ingen av de elektroniska versionerna är fritt tillgänglig.

11 Se <http://lingweb.eva.mpg.de/ids/> och <http://world.livingsources.org/semanticfield/>.

tillgängliga listor är dels goda kärnvokabulärkandidater, dels tillhandahåller de en koppling till denna kärnvokabular i många andra (och i det här sammanhanget ovanliga) språk. Större delen av listorna har försetts med SALDO-betydelseidentifierare.[12]

Svenska Wiktionary innehåller ungefär 43.000 ingångar,[13] uppdelade i betydelser med definitioner. Definitioner är sällsynta eller obefintliga i andra fria lexikonresurser.

3. Hopkoppling av resurserna

Förhoppningsvis har det framgått av ovanstående korta beskrivningar att de befintliga resurserna är mycket heterogena med avseende på sitt innehåll. De är lika mångskiftande ifråga om lagringsformatet, som varierar från tabbseparerade textfiler till flera olika SGML- och XML-format. Denna variation är egentligen inget att förvånas över, eftersom resurserna har utvecklats för olika ändamål av olika grupper av forskare, såväl lingvister som språkteknologer, och till och med mannen på gatan (Synlex).

En huvuduppgift i SweFN++-projektet blir således att harmonisera innehållen i dessa resurser och även att säkerställa att de kan användas som lexikalisk komponent i befintliga språkteknologiska verktyg. Vi behöver också utarbeta strategier för att hantera det faktum att vissa typer av information kommer att vara ojämnt fördelade i den integrerade resursen. T.ex. kommer information om syntaktisk valens att finnas för ungefär en fjärdedel av ingångarna i den integrerade resursen. Å ena sidan vill man kunna använda denna information när man har den, men å den andra vill man inte vara beroende av att den finns, eftersom den saknas i majoriteten av fallen. Sedan är det även en intressant metodologisk fråga

12 Se <http://spraakbanken.gu.se/swefn/resurser/lwt-meanings.html>.

13 Se <http://sv.wiktionary.org/>.

i vilken mån man kan lägga till sådan information för ingångar som saknar den genom att utnyttja annan information som redan finns i resursen, t.ex. om sammansättningar eller semantisk typ.

Detta harmoniserings- och standardiseringsarbete bedriver vi redan helt oberoende av SweFN++, bland annat inom det europeiska samarbetet CLARIN, som har som mål att få till stånd en europeisk infrastruktur för språkresurser.[14]

Harmoniseringen av de befintliga resurserna har två aspekter: *dataformat* och *informationsmodell*. Dataformatet handlar om hur informationen lagras i filer eller databaser, t.ex. i form av XML-dokument av en viss typ. Dataformatet är förvisso mycket viktigt för den praktiska hanteringen av resurserna, men det är inte i grunden svårhanterat. Det kan i stor utsträckning hanteras automatiskt med datorprogram, så det kommer vi att ta itu med senare. Det finns numera en ISO-standard för lexikonresurser (LMF 2008), som vi i någon form kommer att anamma för vår integrerade resurs.

Informationsmodellen kan man däremot behöva arbeta mycket med. En förutsättning för att man ska kunna koppla ihop resurserna är att de har åtminstone någon informationskategori gemensam (det behöver inte vara samma kategori för alla resurserna; det räcker i princip att man kan koppla ihop dem parvis) och att den kategorins grundstruktur sammanfaller eller kan fås att sammanfalla mellan resurserna.

Att detta inte är ett trivialt problem illustreras bäst med ett konkret fall. Atwell et al. (2000) redogör för ett projekt med det till synes okomplicerade målet att harmonisera nio olika engelska ordklasstagguppsättningar. Efter många experiment med olika metoder för att automatiskt konvertera mellan tagguppsättningar kom man till slut fram till att detta var omöjligt på grund av tagguppsättningarnas olika struktur. Det mest effektiva var istället att

14 Se <http://www.clarin.eu>.

helt enkelt ta bort de ursprungliga taggarna och tagga om texten med en annan tagguppsättning.

I det fallet kunde man stödja sig på en beprövad metod som innebär att en ordklasstaggare tränas på ett korrekt taggat korpusmaterial. I vårt fall finns inte den möjligheten. Det finns ingen känd metod för att automatiskt strukturera ett lexikon i lämpliga semantiska eller formella enheter. Man kan alltså förvänta sig att den manuella arbetsinsatsen när det gäller att integrera de befintliga resurserna huvudsakligen kommer att bestå i att definiera kopplingen mellan dem. Då är en viktig fråga om samma sak gäller här som i fallet ordklasstaggning, alltså att man inte skulle hitta många fall av ett-till-ett-avbildning mellan de enheter och informationskategorier i lexikonresurserna som man är intresserad av.

3.1. Länkning via betydelse-ID

De enheter vi vill använda för integreringen av våra resurser är *betydelser*. Det är uppenbart att betydelser är centrala i nästan alla slags lexikonresurser; även de resurser där språklig form står i förgrunden bygger ytterst på att lexikonsammanställaren har tagit ställning till ords och andra lexikonenheters betydelser. I en kontext där resurserna ska användas för automatiskt processande, är det viktigt att vi har ett formellt väldefinierat explicit och entydigt sätt att referera till betydelser. Därför bildar SALDO kärnan i den integrerade resursen.

SALDO är som nämnts ett betydelselexikon och det har designats för att kunna användas i automatisk språkbearbetning. Alla identifierare i SALDO har därför vissa formella egenskaper gemensamma. I SALDO definieras fyra sorters lexikala objekt (exempel inom parentes): *betydelser* (grad..1), *lemgram* (grad.. nn.1), *ordklasser* (nn) och *böjningsparadigm* (nn_3u_film). Ordklasser och böjningsparadigm är i princip slutna mängder (även om de förändras över tid under arbetet med SALDO). Betydelser

och lemgram motsvarar grovt sett språkligt innehåll och språkligt uttryck på det lexikala planet. Identifierarna har en formell syntax vars yttre ram är att de måste vara giltiga XML-namn (XML 2008), därför att vi utan hinder vill kunna använda dem i de formalismer som nu utvecklas för den semantiska webben (t.ex. RDF och OWL) och som har börjat spela en viktig roll i språkteknologisammanhang, en roll som bara förväntas öka i betydelse över de närmaste åren. Detaljerna i identifierarnas formella syntax tar bl.a. hänsyn till att det är enklare för människor att arbeta med representationer som bär någon relation till det som representeras, än med (ur mänsklig synvinkel) arbiträra koder. Därför kan man t.ex. av paradigmidentifieraren ovan utläsa att den gäller utrala (u_) substantiv (nn_) av tredje deklinationen (_3) som böjs och i sammansättningar beter sig som ordet *film* (_film). Slutligen är identifierarna unika, vilket betyder att inget annat ska behöva användas än dessa för att referera till ett objekt i lexikonbeskrivningen, t.ex. (genererade) databasnycklar.

SALDOs betydelseidentifierare (grad..1 – grad..9 för grundformen *grad*) är avsiktligt utformade för att inte avspegla hierarkiska eller andra relationer mellan betydelser. All uppdelning i huvud- och underbetydelser liksom synonymi och andra lexikala betydelserelationer måste uttryckas explicit, separat från betydelserna själva. På det viset gör SALDOs betydelseidentifierarsystem enbart det minimala antagandet att vi kan urskilja separata lexikala betydelser, men utan att ta ställning till hur dessa betydelser ska relateras till varandra, något som erfarenhetsmässigt är både besvärligt och kontroversiellt. Med denna lösning kan vi i princip tillåta ett godtyckligt antal alternativa semantiska strukturer för en viss delmängd av lexikonbetydelser eller hela lexikonet.

Lemgram är vår term för den grundläggande formenheten i SALDO.[15] Den definieras genom en grundform och en uppsätt-

15 I princip kan lemgrammet ses som en generalisering av lemmat i Alléns lemma-lexemmodell (Allén 1967). I praktiken har existensen av den väl

ning formella egenskaper, noga räknat ordklass, böjningsmönster och sammansättningsform. Ordklassuppsättningen bygger på det traditionella systemet, men är betydligt mer differentierat, med tillägg för bl.a. flerordsenheter. F.n. finns 37 ordklassbeteckningar i SALDO. Böjningsmönster och sammansättningsform ger tillsammans de paradigm som identifieras i SALDO, just nu 1130 stycken.[16] En hel del av mångfalden förklaras av att paradigmen ska fånga sammansättningsbeteendet hos lemgrammen. Exempelvis har tredje deklinationens utrala substantiv (de som slutar på konsonant i singular och lägger till *-er* i plural) fyra grundparadigm, som enbart skiljer sig åt med avseende på hur de tar sammansättningsfogen *-s-*: inte som första led och optionellt i andra positioner (nn_3u_film), optionellt i alla positioner (nn_3u_tid), inte i någon position (nn_3u_karbid) samt obligatoriskt i alla positioner (nn_3u_salong).

Eftersom de andra lexikonresurserna är orienterade mot innehåll eller uttryck eller båda och eftersom SALDO är vår mest omfattande resurs och den som är mest konsekvent utformad för användning i språkteknologitillämpningar, blir det naturligt att använda SALDOs betydelse- och lemgramidentifierare för att koppla ihop resurserna med varandra. Alla resurser kopplas således till SALDO och via SALDO till andra resurser.

Den stora frågan blir då som vi såg ovan hur mycket manuellt arbete detta kan tänkas innebära och hur mycket som skulle kunna utföras rent automatiskt. Det är här Zipf kommer in i bilden.

---

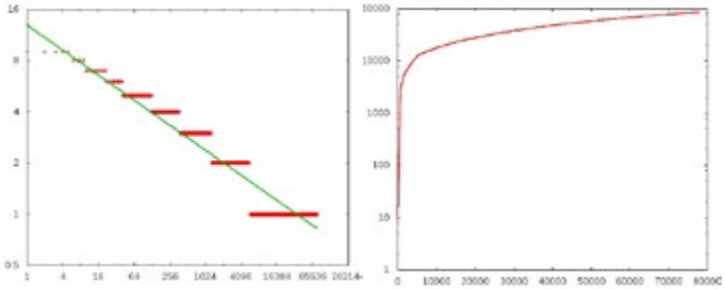
inarbetade engelska termen *lemma* 'grundform' visat sig leda till oönskad begreppsförvirring när man försöker använda samma term för denna formellt definierade enhet (som till råga på allt ligger nära det som på engelska brukar kallas *lexeme*). Därför har vi valt en helt ny term för den; det vi här kallar lemgram och betydelse (och på engelska *lemgram* och *sense*) motsvarar alltså ganska väl lemma och lexem i Alléns modell (även om de metodologiska och ontologiska grundvalarna skiljer sig åt; se Borin 2008).

16 Se <http://spraakbanken.gu.se/swe/forskning/saldo/statistik>.

4. Zipfs lag och lexikonbetydelser

Lingvisten George Kingsley Zipf (1902–1950) har gett namn åt en av de mest kända språkliga statistiska lagbundenheterna, *Zipfs lag* (Zipf 1935). Han upptäckte att ordfrekvenserna i en textkorpus sjunker exponentiellt. Om det oftast förekommande textordet i korpusen har frekvensen *M* och om man ställer upp korpusens ord rangordnade efter sjunkande frekvens, kommer man att finna att ordet på plats *n* har ungefär frekvensen *M/n*, d.v.s. korpusens vanligaste ord förekommer sju gånger så ofta som ordet på plats 7. Det här är en idealisering; i en riktig korpus kommer man oundvikligen så småningom ner till frekvensen 1, som typiskt omfattar ungefär hälften av alla ord i korpusen. Samma sak gäller även för fördelningen av andra typer av språkliga enheter i text.

Om vi undersöker grundformer i SALDO med avseende på hur många betydelser de uttrycker kan vi till att börja med konstatera att den ena extremen är nio betydelser (grundformerna *grad* och *rå*) och den andra naturligtvis en betydelse. De två kurvorna i figur 1 visar att fördelningen av betydelser över grundformer uppvisar ett Zipfbeteende. Om man lägger ut rangordning och antal betydelser (per grundform) i ett koordinatsystem med logaritmisk skala på båda axlarna, ska man kunna approximera dem med en rak linje som sluttar ner åt höger (kurva a; punkterna – som då det finns många grundformer med samma antal betydelser smälter samman till tjocka vågräta streck – visar antal betydelser och den heldragna linjen visar den idealiska Zipffördelning som bäst passar till den faktiska fördelningen av betydelser över grundformer). Om man plottar hur antalet betydelser växer allteftersom man lägger till enheter med sjunkande frekvens, ska man få en kurva som stiger brant och sen snabbt planar ut, som i kurva b (med logaritmisk y-axel).

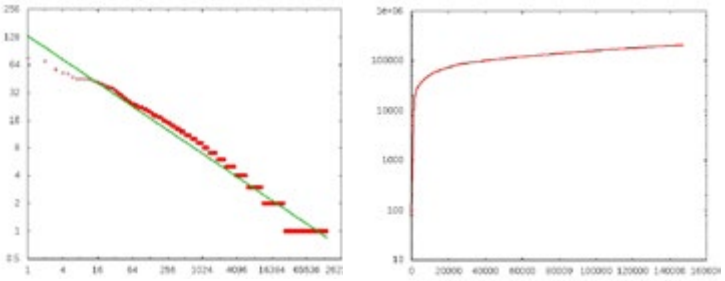


a: rangordning – antal betydelser b: rangordning – betydelsetillväxt

Figur 1: Zipfbeteende i SALDO

Vårt antagande är helt enkelt att detta är ett beteende som inte är specifikt för SALDO, utan något som kännetecknar ordförrådet i ett språk generellt, och därmed alla slags lexikaliska resurser, vilket har stor betydelse då resurser ska kopplas ihop (se nedan). Man kunde å andra sidan tänka sig att detta inte är ett generellt beteende utan något som beror på att SALDO i genomsnitt har relativt få betydelser per grundform. Ett naturligt sätt att pröva antagandet skulle då vara att finna en lexikonresurs som är så olik SALDO som möjligt i det avseendet och undersöka om den beter sig på samma sätt. Det engelska Princeton WordNet (PWN; Fellbaum 1998)[17] är en sådan resurs. Det påpekas ofta om PWN i litteraturen hur (överdrivet) fin dess betydelseindelning är. De mest flertydiga grundformerna i PWN har en storleksordning fler betydelser än i SALDO. De tre toppositionerna i PWN upptas av *break* med 75, *cut* med 70 och *run* med 57 betydelser. När man grafiskt visar hur antalet betydelser per grundform fördelar sig över hela PWN på samma sätt som ovan för SALDO, blir bilden dock mycket likartad (figur 2). Vi får kurvor av samma form, bara förskjutna i höjdled. I SALDO täcker andelen med en enda betydelse per grundform 93 % av grundformerna (86 % av betydelserna), medan den i PWN täcker 81 % av grundformerna (58 % av betydelserna).

17 Se <http://wordnet.princeton.edu/>. För det experiment som beskrivs nedan har Princeton WordNet version 3.0 använts.



a: rangordning – antal betydelser b: rangordning – betydelsetillväxt

Figur 2: Zipfbeteende i Princeton WordNet 3.0

Omvänt betyder ju detta att endast 7 % av grundformerna i SALDO (och 19 % av grundformerna i PWN) är flertydiga. I vårt arbete med att koppla ihop SALDO med de andra resurserna har vi empiriskt iakttagit samma sak. Konsekvenserna av detta är metodologiskt högst löftesrika: Eftersom de flesta grundformerna bara bär en betydelse i våra lexikonresurser, kan vi för majoriteten av betydelserna i lexikonresurserna reducera problemet att jämföra betydelser med varandra till det mycket enklare problemet att para ihop grundformer. Således kan vi huvudsakligen automatiskt – plus en begränsad manuell insats – koppla ihop lexikonresurserna och få acceptabel precision för praktiska tillämpningar.

5. Mot en integrerad lexikonresurs för svensk språkteknologi

För att snabbt komma igång med SweFN++-projektet har vi valt ett arbetssätt som innebär återanvändning inte bara av de lexikonresurser som står i fokus i projektet, utan även av programvara och standardverktyg. I den pilotfas vi nu befinner oss av projektet använder vi oss av vad man skulle kunna kalla för ”barfotalösningar”. Istället för att först med en stor arbetsinsats försöka bygga

en integrerad mjukvarulösning för ett problem vars fullständiga konturer vi inte känner än, har vi kopplat ihop befintliga verktyg och komponenter med hjälp av relativt lättviktiga webbtjänster och små specialprogram, där alla inblandade lägger resultatet av sitt arbete med de olika ingående resurserna i ett gemensamt centralt datalager. Där sker formella kontroller av data och en del andra bearbetningar automatiskt med regelbundna intervaller. Bland annat läggs nya versioner av resurserna, automatgenererad statistik och felrapporter upp på Språkbankens webbplats minst en gång per dygn.[18] En i vårt tycke stor fördel med detta sätt att organisera arbetet är *transparens*; projektwebbsidorna där denna information läggs upp är synliga för alla och vi tar gärna emot synpunkter på alla aspekter av projektet.

Genom att på detta vis koppla olika komponenter och verktyg löst till varandra kan vi med relativt små arbetsinsatser experimentera oss fram emot ett fungerande arbetsflöde för hela projektet, där vi förhoppningsvis metodologiskt ska kunna finna en optimal kombination av automatiska metoder och manuellt arbete.

Mer specifikt kommer vi att under den närmaste framtiden utforska bl.a. följande metodologiska aspekter som främst har att göra med integreringen av de befintliga resurserna (för de aspekter som närmast berör utbyggnaden av frasnätskomponenten, se Borin et al. 2010):

Hur kan vi använda existerande information i resurserna för att automatiskt tillföra saknad information? Kan vi t.ex. anta att en sammansättning är av samma semantiska typ som sitt slutled för att utvidga informationen från SIMPLE-lexikonet till betydelser som inte finns i det (vilket är omkring 90 % av betydelserna i hela resursen)? Kan vi koppla ord i Synlex till flertydiga grundformer i SALDO genom att jämföra deras semantiska närkontext i

18 Se <http://spraakbanken.gu.se/swe/swefn> och <http://spraakbanken.gu.se/saldo/>.

SALDO med de angivna synonymerna i Synlex, för att på det viset välja rätt bland alternativen?

Kan vi använda oss av korpusverktyg, t.ex. en parser, och utifrån ords syntaktiska kontext i korpusar – t.ex. objekt till ett visst verb eller en viss (semantisk) klass av verb – plus deras semantiska egenskaper hel- eller halvautomatiskt komplettera vår resurs med syntaktisk valens för sådana lemgram som inte finns i PAROLE-lexikonet?

Hur kan vi utforma en användarmiljö där flera personer kan arbeta samtidigt med olika delresurser – t.ex. SALDO-, Swesaurus- och SweFN-komponenten av SweFN++ – men där vi ändå kan säkerställa att resurserna hålls synkroniserade? Idag har vi ett enkelt diagnostiskt program och en webbsida som visar ifall några beroenden gentemot SALDO har brutits efter en uppdatering av någon av resurserna. Webbsidan uppdateras nu en gång per dygn,[19] men när man arbetar aktivt med en resurs önskar man sig naturligtvis en mer direkt återkoppling på det man gör. Det är viktigt att understryka att behovet av en sån här funktion har blivit riktigt uppenbart bara efter det att den faktiskt har blivit verklighet, om än i väldigt enkel form.

Alla dessa frågor och många andra hoppas vi kunna utforska med hjälp av de tillgångar som vi har i form av lexikonresurser, korpusar och verktyg för språklig uppmärkning av korpusar, för att så småningom kunna erbjuda svensk språkteknologi en högvärdig, framtidssäker och fritt tillgänglig lexikonresurs i form av SweFN++.

Litteratur

Allén, Sture 1967. *Studier över nusvenskans vokabulärsystem. Opublicerad rapport*. Institutionen för nordiska språk, Göteborgs universitet.

19 Se <http://spraakbanken.gu.se/swe/forskning/swefn/beroendeanalys>.

Atwell, Eric, George Demetriou, John Hughes, Amanda Schiffrin, Clive Souter & Sean Wilcock 2000. Comparing linguistic interpretation schemes for English corpora. I: *Proceedings of COLING LINC-2000 Workshop on Linguistically Interpreted Corpora*. Luxembourg: ACL. 1–10.

Boas, Hans C. (utg.) 2009. *Multilingual framenets in computational lexicography*. Berlin: Mouton de Gruyter.

Borin, Lars 2005. Mannen är faderns mormor: *Svenskt associationslexikon reinkarnerat*. I: *LexicoNordica* 12: 39–54.

Borin, Lars 2008. Lemma, lexem eller mittemellan? Ontologisk ångest i den digitala domänen. *Nog ordat? Festskrift till Sven-Göran Malmgren*. Göteborgs universitet, Meijerbergs arkiv för svensk ordforskning. 59–67.

Borin, Lars, Dana Dannélls, Markus Forsberg, Dimitrios Kokkinakis & Maria Toporowska Gronostaj 2010. The past meets the present in Swedish FrameNet++. I: *Proceedings of Euralex 2010*.

Borin, Lars & Markus Forsberg 2009a. All in the family: A comparison of SALDO and WordNet. I: *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. NEALT Proceedings Series, 7. 7–12.

Borin, Lars & Markus Forsberg 2009b. Something old, something new: A computational morphological description of Old Swedish. I: *LREC 2008 workshop on language technology for cultural heritage data (LaTeCH 2008)*. Marrakech: ELRA. 9–16.

Borin, Lars & Markus Forsberg 2010. From the People's Synonym Dictionary to fuzzy synsets – first steps. I: *Proceedings of the LREC 2010 workshop Semantic relations. Theory and Applications*. Valletta: ELRA.

Borin, Lars, Markus Forsberg & Dimitrios Kokkinakis 2010. Diabase: Towards a diachronic BLARK in support of historical studies. I: *Proceedings of LREC 2010*. Valletta: ELRA.

Borin, Lars, Markus Forsberg & Lennart Lönngren 2008. The hunting of the BLARK – SALDO, a freely available lexical database

for Swedish language technology. I: J. Nivre, M. Dahllöf & B. Megyesi (utg.), *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*. Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7. 21–32.

Bresnan, Joan (utg.) 1982. *The mental representation of grammatical relations*. Cambridge, Massachusetts: MIT Press.

Bresnan, Joan & Ronald Kaplan 1982. Lexical-Functional Grammar: A formal system for grammatical representation. I: J. Bresnan (utg.), *The mental representation of grammatical relations*. Cambridge, Massachusetts: MIT Press. 173–281.

Bring, Sven Casper 1930. *Svenskt ordförråd ordnat i begreppsklasser*. Stockholm: Hugo Gebers Förlag.

Dalin, Anders Fredrik 1850–53. *Ordbok öfver svenska språket. Vol. I–II*. Stockholm.

Fellbaum, Christiane (utg.) 1998. *WordNet: An electronic lexical database*. Cambridge, Massachusetts: MIT Press.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum & Ivan Sag 1985. *Generalized phrase structure grammar*. Oxford: Basil Blackwell.

Johnson, Christopher & Charles Fillmore 2000. The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. I: *Proceedings of the first meeting of the NAACL*. Seattle: ACL. 56–62.

Järborg, Jerker 2001. *Roller i Semantisk databas* (Research Reports from the Department of Swedish, No. GU-ISS-01-3). University of Gothenburg: Dept. of Swedish Language.

Kann Viggo & Magnus Rosell 2006. Free construction of a free Swedish dictionary of synonyms. I: *Proceedings of the 15th NODALIDA*. Dept. of Linguistics, University of Joensuu. 105–110.

Koptjevskaja-Tamm, Maria, Martine Vanhove & Peter Koch 2007. Typological approaches to lexical semantics. *Linguistic Typology* 11(1): 159–186.

Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas & An-

tonio Zampolli 2000. SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Lexicography* 13(4): 249–263.

LMF 2008. *Language resource management: Lexical markup framework*. International standard ISO 24613:2008. First edition, 2008-11-15. <http://www.lexicalmarkupframework.org/>

Lönngren, Lennart 1989. A Swedish associative thesaurus. I: *Euralex 1998 Proceedings*. Liège: University of Liège. 467–474.

Pollard, Carl & Ivan Sag 1994. *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Saddock, Jerrold 1991. *Autolexical syntax*. Chicago: University of Chicago Press.

Schlyter, Carl Johan 1887. *Ordbok till samlingen av Sveriges gamla lagar*. Saml. av Sveriges gamla lagar 13. Lund.

Small, Steven L. & Chuck Rieger 1982. Parsing and comprehending with word experts (a theory and its realization). I: W.G. Lehnert & M.H. Ringle (utg.): *Strategies for natural language processing*. Hillsdale, NJ: L. Erlbaum, 89–147.

Söderwall, Knut Fredrik 1884. *Ordbok öfver svenska medeltidsspråket*. Vol. I–III. Lund.

Söderwall, Knut Fredrik 1884. *Ordbok öfver svenska medeltidsspråket. Supplement. Vol. IV–V*. Lund.

XML 2008. Extensible Markup Language (XML) 1.0 (Fifth Edition). W3C Recommendation 26 November 2008. <http://www.w3.org/TR/xml/>

Zipf, George K. 1935. *The psycho-biology of language*. Boston: Houghton Mifflin.

Lars Borin
professor
Språkbanken
Institutionen för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
lars.borin@svenska.gu.se

Bruk av et norsk leksikon til tagging og andre språkteknologiske formål

Kristin Hagen & Anders Nøklestad

Norsk ordbank (the Norwegian Word Bank) is an electronic lexicon for the two Norwegian written standards, *Bokmål* and *Nynorsk*. It forms the basis of many, probably most, of the existing language technology tools for Norwegian. The lexicon is based on the entries and inflectional information found in the dictionaries *Bokmålsordboka* and *Nynorskordboka* as well as word lists and inflectional patterns developed by IBM Norway. We present some background information about the lexicon and show how it has been applied to a variety of language technology tools and various applications for end users. Since the lexicon was developed from resources meant for use by human readers, much work has been devoted to modifying the lexicon to make it better suited for use in language technology, and the main focus of our paper is on this work.

1. Innledning

Norsk ordbank er et elektronisk leksikon for de to norske skriftspråksstandardene bokmål og nynorsk. Innholdet i Ordbanken kommer dels fra håndordbøkene *Bokmålsordboka* og *Nynorskordboka*, og dels fra elektroniske ordlister utviklet ved IBM Norge. Ordbanken ble laget i 1996 i forbindelse med Taggerprosjektet, et prosjekt som skulle utvikle en morfologisk og syntaktisk tagger for bokmål og nynorsk.

Mye av bakgrunnsmaterialet til Ordbanken stammer altså fra ordbøker, skrevet og redigert for å tilfredsstille menneskelige lesere i et trykt medium. I den første delen av denne artikkelen vil vi redegjøre for opprinnelsen til Ordbanken med fokus på hvilke endringer og tilpasninger som måtte gjøres for å få innholdet best

mulig egnet som elektronisk ordbok brukt til morfologisk og syntaktisk tagging.

Norsk ordbank er et enkelt fullformsleksikon som mangler semantisk tilleggsinformasjon som ville ha vært nyttig i en del sammenhenger. Etter Taggerprosjektets avslutning har Ordbanken likevel vært brukt i mange ulike språkteknologiske applikasjoner og verktøy. Den siste delen av artikkelen vil gå nærmere inn på dette og beskrive noen av verktøyene i større detalj.

2. Kort om bakgrunnen til Norsk ordbank

Norsk ordbank, eller *taggerbasen* som leksikonet ble kalt i begynnelsen, ble opprettet i 1996 i regi av *Taggerprosjektet*.[1] Formålet med dette prosjektet var å utvikle en disambiguerende morfologisk og syntaktisk tagger for norsk,[2] og for å komme i gang med dette arbeidet trengtes det elektroniske fullformsordlister. Ved prosjektstart fantes det ikke andre lister tilgjengelig enn IBM-ordlistene som *Dokumentasjonsprosjektet* hadde kjøpt fra *IBM Norge*.[3] For å komme i gang med utviklingen av taggeren ble det derfor beslut-

1 Taggerprosjektet (1996–1999, ledet av Janne Bondi Johannessen) ble beregnet til seks årsverk. Tre av disse ble finansiert fra Norges forskningsråd, to fra Dokumentasjonsprosjektet (ledet av Christian-Emil Smith Ore) og ett fra Tekstlaboratoriet (ledet av J. B. Johannessen). I tillegg kom grunnmateriale fra *Bokmålsordboka* og *Nynorskordboka* utviklet ved seksjon for Leksikografi ved Universitetet i Oslo, IBM-ordlistene med grammatiske koder for bokmål og nynorsk, programvare for disambigueringsdelen av taggeren fra Lingsoft og argumentstruktur for verb fra NorKompLeks, NTNU.

2 Taggeren som ble utviklet, heter i dag Oslo-Bergen-taggeren, se referanse.

3 Samtidig med Taggerprosjektet (1996) fikk NorKompLeks-prosjektet ved NTNU støtte fra Norges forskningsråd til å lage et maskinleselig språkteknologisk leksikon for norsk (bokmål og nynorsk). Dette arbeidet kunne Taggerprosjektet dessverre ikke benytte seg av, siden prosjektet var avhengig av elektroniske ordlister fra prosjektstart for å utvikle taggeren.

tet å bruke disse ordlistene sammen med oppslagsord og grammatiske opplysninger fra *Bokmålsordboka* og *Nynorskordboka* for å lage en ordbase med fullformer for taggeren. Dette arbeidet blir beskrevet i større detalj i kapittel 3.

Flere miljøer ved Universitetet i Oslo (UiO) var involvert i utviklingen av taggerbasen: Dokumentasjonsprosjektet, Seksjon for leksikografi og målføregransking ved Institutt for nordistikk og litteraturvitenskap og Tekstlaboratoriet ved Institutt for lingvistiske fag. For å sikre basens videre utvikling, vedlikehold og drift fikk taggerbasen i 2001 et styre med overordnet ansvar for dette. Styret har medlemmer fra de ulike UiO-miljøene samt Språkrådet. Basen ble omdøpt til *Norsk ordbank* og eies i dag av Institutt for lingvistiske og nordiske studier (ILN), der alle miljøene nevnt ovenfor nå er samorganisert, og Språkrådet.

Ordbanken blir finansiert ved hjelp av egeninnsats fra ILN og Språkrådet samt med enkelte midler fra salg. Ordbanken er for øvrig nedlastbar på GPL-lisens.[4]

Det har bare vært gjort mindre endringer i Ordbanken siden oppstarten i 1996. LOGON-prosjektet (Oepen et al. 2007) finansierte noe arbeid for bokmål i 2001, og Språkrådet finansierte oppdateringer etter nyere Språkråds-vedtak i 2007, samt innlegging av nyord fra Bokmålsordboka og Nynorskordboka.

Ordbanken inneholder i dag 151 229 lemmaer for bokmål og 126 323 lemmaer for nynorsk.

3. Utviklingen av et elektronisk fullformsleksikon for norsk

Utviklingen av Ordbanken var fra starten av styrt av behovet til taggeren som skulle utvikles. For å gjenkjenne alle ord i en gitt

4 GNU General Public License (GPL), jf. http://www.gnu.org/licenses/gpl.html

tekst trengte taggeren fullformsordlister, det vil si ordlister med lemmaer (for eksempel *bil*) og alle lemmaets mulige bøyingsformer (for eksempel *bil bilen biler bilene*). Det var i tillegg ønskelig å lagre lemmaene og deres bøyingsformer på en måte som gjorde at de var lette å finne igjen og lette å redigere om det skulle vedtas endringer i ortografi eller bøyingsmåte senere. Det var også ønskelig at bøyingskodene skulle følge ordklasseterminologien i *Norsk referansegrammatikk (*Faarlund/Lie/Vannebo 1997) med for eksempel *determinativer* i stedet for *artikler*, og *subjunksjoner* i stedet for *underordnende konjunksjoner.*

I kapitlene nedenfor vil vi gå igjennom de ulike kildene til Ordbanken og beskrive nærmere hva som ble gjort.

3.1. IBM-ordlistene

Listene fra IBM (Engh 1994) inneholdt nærmere 122 000 lemmaer for bokmål og mer enn 110 000 for nynorsk. Lemmaene var hentet fra både korpus og ordbøker, og alle ord var utstyrt med en bøyingskode slik at fullformene kunne avledes, se eksempel 1 på neste side.

Listene ble utviklet for å brukes i IBMs språkteknologiske produkter, for eksempel i IBMs stavekontroll. Når listene nå skulle brukes til tagging, hadde de en del åpenbare svakheter:

1. Listene inneholdt ikke sideformer.[5]
2. Listene var normative.
3. Bøyingsmønstrene var komplekse og krevende å vedlikeholde.

En tagger bør kunne gjenkjenne alle ordene i en tekst, også sideformer og de mest vanlige feilskrevne ord og bøyinger. IBM-listene måtte derfor suppleres, både med hensyn til lemmautvalg og med hensyn til bøyingsinformasjon, se avsnitt 3.2. og 3.3.

5 *Sideform* er definert slik i *Bokmålsordboka*: "ord- el. bøyningsform i offisiell rettskrivning som ikke kan brukes i lærebøker og i sentraladministrasjonen, t forskj fra *hovedform*".

```
#B_N10075.010
*subst fem
**fullst

Ø900 Ø  NOB_FN
01
02 a,en
03 er
04 ene
05 s
06 as,ens
07 ers
08 enes
```

Eksempel 1: Bøyingskode for bokmålssubstantiv som for eksempel *dør*. Ved hjelp av bøyingsmønsteret kan fullformene *dør, døra* eller *døren, dører, dørene* avledes sammen med tilhørende genitivsformer *dørs, døras* eller *dørens, dørers, dørenes.*

Bøyingsmønstrene fra IBM måtte også endres av samme årsak. Fordi IBM-listene skulle brukes normativt, var det lagt mye arbeid i å normere bruken av flertall for substantiv, komparative former for adjektiv og adjektivavledninger av verb. Dette arbeidet ble gjort i samarbeid med Språkrådet. For taggeren, derimot, var denne normeringen mer til skade enn til gagn. Flertallssubstantiv som *adganger* og *afasier* er ikke vanskelig å finne ved søk i korpus, selv om ordene har en bøyingskode som utelukker flertallsformer i IBM-listene. Og igjen, uansett hva man måtte mene om flertallsformer av slike substantiv, må taggeren kunne tagge eksempler som ”Dette gjelder også for *adganger* som er gitt til grupper” eller ”en gruppe språkforstyrrelser som kalles *afasier*”. Derfor ble alle substantiv – med unntak av ubøyelige substantiv som *gøy* – gitt flertallsformer. Alle adjektiver ble også gitt komparative former

og alle verb fikk adjektivavledninger slik at eksempler som "... det generelt er en myte at noen språk liksom er *ordfattigere* enn andre" og "Jeg syntes de lager for *brummete* lyder" kunne få en analyse av taggeren. (*Ordfattig* hadde opprinnelig kode uten komparative former, og *brumme* hadde kode uten adjektivavledning.)

Som eksempel 1 ovenfor viser, var bøyingskodene i IBM-listene komplekse og uoversiktlige med flere mulige bøyinger for hver kode. Dette ble løst opp slik at systemet skulle bli lettere å forstå og vedlikeholde. I Ordbanken ble f.eks. koden for ordet *dør* fra eksempel 1 løst opp, slik at ordet nå har to bøyingskoder, 700 og 900, som genererer de samme fullformene som tidligere,[6] se eksempel 2. Legg merke til at hvert bøyingssuffiks er unikt definert med et tall som angir grammatisk informasjon for dette suffikset (f.eks. angir kombinasjonen 700, 02 bestemt form entall av substantiv).

700	900
----	----
01	01
02 en	02 a
03 er	03 er
04 ene	04 ene

Eksempel 2: Bøyingskoder for bokmålssubstantiv som for eksempel *dør*. Ved hjelp av bøyingsmønstrene 700 og 900 kan fullformene *dør, døra* eller *døren, dører, dørene* avledes.

IBM-listene inneholdt ikke bare de vanligste lemmaene, men også en god del egennavn og mange kreative sammensetninger av typen *bruksvakthold* og *prisdepartement*. Slike sammensetninger er kanskje

6 I Ordbanken er genitivs-*s* ikke med i bøyingskodene. Oslo-Bergen-taggeren har en egen modul som forsøker å avgjøre om en apostrof eller en *s* markerer gentitiv. Grunnen til dette er at genitiv ikke er en bøyingskategori ved norske substantiver, men at den såkalte genitivs-*s* i stedet er et klitikon som kan hektes på en frase: "jenta som roptes (genitivs-*s*) hatt".

ikke så frekvente, men gjør heller ingen skade for taggeren. Oslo-Bergen-taggeren har for øvrig en egen sammensetningsmodul som analyserer nylagede sammensetninger, slik at en ikke er avhengig av en fullformsordliste med alle mulige tenkelige sammensetninger.

3.2. Innhold fra *Bokmålsordboka* og *Nynorskordboka*

Fra *Bokmålsordboka* og *Nynorskordboka* fikk Ordbanken oppslagsord med grammatiske opplysninger og normeringsinformasjon. Lemmatilfanget i ordbøkene var bare delvis overlappende med lemmatilfanget i IBM-ordlistene, i tillegg til at ordbøkene naturligvis inneholdt alle sideformer. Med ordbøkene ble Ordbanken altså styrket med flere lemmaer og mer informasjon om hvert lemma, men også her oppstod det problemer i forhold til taggerens behov:

1. De grammatiske opplysningene fra ordbøkene var lite spesifikke.
2. Ordbøkene hadde flertydige faste uttrykk som oppslagsord.
3. Ordbøkene inneholdt mange lavfrekvente oppslag med svært frekvente homonymer.

Ordbøkenes bøyingskoder, for eksempel *m1*, er ikke spesifikke nok til å kunne konverteres automatisk til bøyingskodesystemet vi endte opp med å bruke i Ordbanken. I Nynorskordboka har for eksempel både *gut* og *låve* bøyingskode *m1*, selv om låve ender på *-e* og dermed trenger en egen bøyingskode *702*. (Med bøyingskode *700* som *gut* har, ville bøyingen for *låven* blitt *låveen*.) For å gi alle ordene i Ordbanken bøyingskoder slik som ordene fra IBM-listene, ble det forsøksvis laget konverteringslister (*m1* → *700*, *f1* → *900*, *v1* → *001* osv.), før lemmaer med bøyingskoder ble korrekturlest manuelt.

For å gjøre ordbøkene mer leservennlige, er faste uttrykk som *av garde* og *til syne* gjengitt som oppslagsord. Dette er oftest en

fordel for taggeren, fordi ordene i uttrykkene hører så fast sammen og bør analyseres sammen. Flertydige faste uttrykk som dette er imidlertid et problem: "Han ville gjøre det *selv om* han ikke fikk lov". Dersom *selv om* er et fast uttrykk, vil taggeren aldri få anledning til å analysere bruken der *selv* og *om* utgjør egne ledd, som i "Hun vasket seg *selv om* kvelden". Uttrykkene fra ordbøkene ble derfor gjennomgått manuelt, og flertydige uttrykk ble merket på en slik måte at de blir utelatt fra taggerens analyse.

For taggeren er det som regel en stor fordel at leksikonet er så rikholdig som mulig. Men når frekvente lemmaer som *kan* (verb), *med* (preposisjon) og *bare* (adverb) har svært lite frekvente homonymer som gjengitt i eksempel 3 nedenfor, skaper det unødig flertydighet for taggeren.

kan: (substantiv) prinsetittel i Mongolia[7]
med: (substantiv) siktemerke; formål, mening
å bare seg: (verb) avholde seg

Eksempel 3: lite frekvente lemmaer som har frekvente homonymer

Gjennom arbeidet med taggeren ble det oppdaget flere slike homonymer, som ble merket slik at de ikke kommer med i taggerens analyse.

Selv om ikke grammatikken fra ordbøkene kunne brukes direkte, er den grammatiske informasjonen tatt inn i Ordbanken sammen med normeringsopplysninger. Etymologi og definisjoner er ikke med, men oppslagsordenes referansenummer er tatt vare på, slik at ordbank og ordbøker kan lenkes sammen.

7 *Kan* og *khan* var sidestilt i *Bokmålsordboka* da Ordbanken ble lagd, men fra 2005 står bare *khan* oppført.

3.3. Tillegg

Mange tekster er dårlig korrekturlest og inneholder feil. Feilstavinger og feilbøyinger hører vanligvis ikke hjemme i ordbaser, men siden Ordbanken ble utviklet for tagging, ble det likevel lagt til en del frekvente feil. Slike tillegg fikk *tillegg* oppført som kilde og *unormert* som normering. Nedenfor følger noen eksempler på feil som ble lagt inn i Ordbanken:

- Frekvente feilstavinger av lemmaer: *almen* i stedet for normert *allmenn* (adjektiv), *arbeide* i stedet for normert *arbeid* (substantiv)
- Frekvente feilbøyinger: *gutta* i stedet for normert *guttene, lederer* i stedet for normert *ledere*

Det ble også lagt inn en del nye lemmaer, forkortelser og uttrykk:

- Nye ord/slang/engelske ord: *drita*, *body*
- Flere egennavn
- Uttrykk: *dann og vann*
- Forkortelser: *adj*

I ordbøkene kan enkelte oppslagsord ha betydninger med ulik ordklasse. *All* er for eksempel oppgitt med ordklasse pronomen i Bokmålsordboka, men i betydningsnummer 5 i ordartikkelen står det likevel at ordet kan være adverb i formen *alt*:

> 5 adv i formen *alt*: helt, allerede *alt fra hun var født* / *han er alt reist* / *alt det hun klager* samme hvor mye hun klager / *alt etter forholdene* i samsvar med / *det er alt etter som en tar det* det kommer an på hvordan ...

Alt er derfor lagt inn som adverb som *tillegg*.

Tilleggene er lagt inn undervegs i forbindelse med arbeidet med taggeren. Alle tillegg stammer fra korpus, men det er ikke systematisk registrert feil og nyord i Ordbanken.

3.4. Argumentstruktur for verb fra NorKompLeks

NorKompLeks-prosjektet (Nordgård 1996) utviklet maskinleselige leksikografiske produkt parallelt med arbeidet med taggeren. Da prosjektet ble avsluttet, ble opplysninger om verbenes argumentstruktur lagt inn i Ordbanken på denne måten:

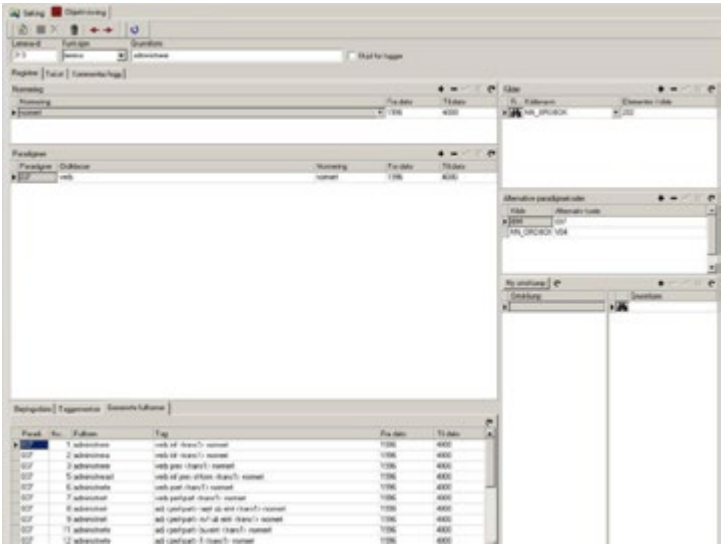
arbeide <intrans1> <predik13>
øke <trans1> <intrans2> <part4/på>

Subkategoriseringsinformasjonen er ikke vedlikeholdt etter prosjektslutt, og for taggerens del viste det seg at informasjonen ikke er fullstendig nok til å kunne brukes i taggeren.

3.5. Database og redigeringsverktøy

EDD (Enhet for digital dokumentasjon[8]) ved ILN, Universitetet i Oslo, er ansvarlig for databasedesign, utvikling og teknisk drift av Ordbanken i dag. Innholdet i Ordbanken er lagret i en Oracle-database, og har et grensesnitt for redigering som også er utviklet av EDD, se figur 1.

8 Enhet for digital dokumentasjon (EDD) ble opprettet for å vedlikeholde og videreutvikle databasene og de elektroniske samlingene fra Dokumentasjonsprosjektet.



Figur 1: Utsnitt av Ordbankens grensesnitt for redigering

4. Eksempler på bruk av Ordbanken

Som nevnt tidligere inneholder Ordbanken relativt enkel informasjon, men denne informasjonen er til gjengjeld svært grunnleggende og nyttig for mange formål, både i forbindelse med språkteknologiske verktøy og for mer brukerrettede applikasjoner. Noen eksempler på kommersiell bruk av Ordbanken er:

- maskinoversettelse
- søkemotorer
- stavekontroll
- ordspill (nettbaserte spill som involverer ord, som f.eks. Scrabble-liknende spill)
- pedagogiske verktøy

Ordbanken har også blitt brukt i mange – trolig de fleste – ikke-kommersielle prosjekter for utvikling av norsk språkteknologi, særlig i kraft av å fungere som leksikon i Oslo-Bergen-taggeren. Noen eksempler på ikke-kommersielle anvendelser av Ordbanken er:

- ordbokssøk på nett[9]
- de fleste norske korpus ved universitetene i Oslo og Bergen
- talespråkstaggere
  - NoTa-taggeren (Nøklestad/Søfteland 2007; Søfteland/Nøklestad 2008)
  - nynorsk dialekttagger (utviklet i forbindelse med prosjektet *Dialektendringsprosessar* ved Universitetet i Bergen)
- maskinoversettelse (LOGON, Apertium)
- analyse av setninger brukt i trebank og grammatikkspill på nett (VISL; Bick 2005)

For ytterligere å demonstrere nytteverdien av Ordbanken, vil vi nå gå litt mer i detalj om to verktøy som nylig har blitt utviklet for bokmål, og som gjør god bruk av informasjonen i Ordbanken: en navnetypegjenkjenner og et anaforløsingssystem. Begge disse verktøyene er beskrevet nærmere i Nøklestad (2009).

4.1. Navnetypegjenkjenning

Navnetypegjenkjenning (eng. *named entity recognition*) går ut på å klassifisere såkalte *named entities*, først og fremst egennavn, i henhold til et sett av kategorier.[10] Nøklestad (2009) beskriver en navnetypegjenkjenner for bokmål som ble utviklet i forbindelse med

9 http://www.bokmålsordboka.uio.no
http://www.nynorskordboka.uio.no

10 Mange navnetypegjenkjennere klassifiserer også datoer, tidsuttrykk, prosenter og pengebeløp. Systemet som er beskrevet her, fokuserer imidlertid på egennavn.

Nomen Nescio-prosjektet (Johannessen et al. 2005), et prosjekt som hadde som formål å lage navnetypegjenkjennere for norsk, svensk og dansk. Prosjektet opererte med følgende navnekategorier:

- Person
- Organisasjon
- Sted
- Hendelse (f.eks. *Statens høstutstilling*, *Kristiansand box cup*)
- Verk (bøker, filmer, musikkalbum o.l.)
- Annet (f.eks. *Bovine-virus*, *Nissan Terrano*)

Systemet som er beskrevet av Nøklestad, bruker såkalte maskinlæringsteknikker for å klassifisere hvert egennavn i en tekst i henhold til disse kategoriene. Maskinlæringsteknikker er teknikker som setter en datamaskin i stand til å lære å utføre en oppgave (f.eks. å klassifisere navn) ved å se på et stort antall riktig klassifiserte eksempler i stedet for at den blir gitt et sett av regler for hvordan oppgaven skal løses. For at en maskin skal kunne lære å klassifisere egennavn, må den få informasjon om hvert navn og omgivelsene navnet står i. Nøklestads system har tilgang til følgende informasjon:

- formen på navnet
  - ”suffikser” (dvs. de tre siste bokstavene i ordet)
  - antall ord i navnet
  - hvorvidt navnet bare inneholder store bokstaver (f.eks. IBM)
- lemmaet til navnet og andre ord i nær kontekst
- ordklassen til ordene i nær kontekst
- syntaktiske relasjoner
- navnelister

Viktige deler av denne informasjonen blir hentet fra Norsk ordbank. Teksten blir disambiguert av Oslo-Bergen-taggeren, som henter lemmaer og ordklasser direkte fra Ordbanken. De syntaktiske relasjonene blir også bestemt av Oslo-Bergen-taggeren, og siden Ordbanken utgjør taggerens leksikon, bidrar den også indirekte til denne typen informasjon. Navnetypegjenkjenneren oppnår et prestasjonsnivå på 83 %, noe som er et godt resultat i internasjonal sammenheng. Dette viser at den enkle informasjonen som finnes i Ordbanken, kombinert med navnelister og ordformer, er tilstrekkelig for å oppnå et godt resultat på denne oppgaven; semantisk informasjon er altså ikke en forutsetning.

4.2. Anaforløsing

Hovedtemaet i Nøklestad (2009) er utviklingen av et system for automatisk anaforløsing (eng. *anaphora resolution*) i bokmål. Anaforløsing, slik denne oppgaven er definert hos Nøklestad, innebærer å finne den nærmeste antecedenten for hver anafor i en tekst.[11] Dette er illustrert i eksempel 4 nedenfor, der oppgaven går ut på å finne den nærmeste antecedenten (om noen) for anaforen *det* i hvert enkelt tilfelle.

Toget traff reinsdyret fordi ...
a) ... *det kjørte for fort*
b) ... *det sto i skinnegangen*
c) ... *det var mørkt*

Eksempel 4: I a) refererer *det* til toget, og *Toget* er derfor nærmeste antecedent; i b) er det *reinsdyret* som er nærmeste antecedent, mens i c) er *det* ikke-referensielt og har derfor ingen antecedent.

11 Nøklestads system er begrenset til å håndtere pronominale anaforer.

I likhet med navnetypegjenkjenneren som ble beskrevet i forrige underkapittel, baserer også anaforløsingssystemet seg på maskinlæring. For å velge riktig antecedent trenger systemet informasjon om anaforen, potensielle antecedenter og forholdet mellom anafor og potensiell antecedent. Nøklestads anaforløsingssystem benytter seg av følgende informasjonskilder:

- avstand mellom anafor og antecedentkandidat
- ordform
- lemma
- ordklasse
- morfosyntaktiske trekk som genus og numerus
- syntaktiske funksjoner
- semantisk informasjon

Igjen kommer vesentlige deler av denne informasjonen fra Norsk ordbank. Lemma, ordklasse og morfosyntaktiske trekk blir hentet direkte fra Ordbanken via Oslo-Bergen-taggeren. Ordklasse er helt essensielt, siden bare nominale ledd blir ansett som potensielle antecedenter i dette systemet. Lemma og morfosyntaktiske trekk (genus og numerus) viser seg også å være blant de aller viktigste informasjonstypene (Nøklestad 2009:216, 226).

Når det gjelder semantisk informasjon, så utgjør navnetypen på antecedentkandidater som er egennavn den klart viktigste faktoren (Nøklestad 2009:217, 226). Navnetypen blir avgjort av navnetypegjenkjenneren beskrevet i forrige underkapittel, og denne navnetypegjenkjenneren benytter seg som nevnt i stor grad av informasjon fra Ordbanken. Som vi også nevnte i forrige underkapittel, bidrar Ordbanken indirekte til disambigueringen av syntaktiske funksjoner, siden den utgjør leksikonet til Oslo-Bergen-taggeren. Vi kan derfor konkludere med at Ordbanken er involvert i de fleste og de viktigste faktorene som systemet bygger på. Dermed spiller den – til tross for sitt relativt enkle innhold – en

avgjørende rolle for å bringe anaforløsingssystemet opp til et prestasjonsnivå på 74,6 %, noe som er et godt resultat i internasjonal sammenheng og signifikant bedre enn det som tidligere har blitt oppnådd for anaforløsing i norsk.

5. Konklusjon

I denne artikkelen har vi vist hvordan leksikonet Norsk ordbank ble til, og hvordan det måtte bearbeides for å kunne brukes til grammatisk tagging og andre språkteknologiske formål. Vi har også gitt noen eksempler på hvordan et relativt enkelt leksikon kan utgjøre grunnlaget for komplekse språkteknologiske verktøy som oppnår gode resultater.

Litteratur

Ordbøker

Bokmålsordboka = Wangensteen, Boye (red.) 2004: *Bokmålsordboka: definisjons- og rettskrivningsordbok*. Oslo: Kunnskapsforlaget.

Nynorskordboka = Anne Engø/Marit Hovdenak/Dagfinn Worren (red.) 2006: *Nynorskordboka. Definisjons- og rettskrivingsordbok*. Oslo: Det Norske Samlaget.

Annen litteratur

Bick, Eckhard 2005: Grammar for Fun: IT-based Grammar Learning with VISL. I: Peter Juel Henrichsen (red.): *CALL for the Nordic Languages*. København: Samfundslitteratur (Copenhagen Studies in Language), 49–64.

Faarlund, Jan Terje/Lie, Svein/Vannebo, Kjell Ivar 1997: *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Johannessen, Janne Bondi/Hagen, Kristin/Haaland, Åsne/Jónsdottir, Andra Björk/Nøklestad, Anders/Kokkinakis, Dimitris/Meurer, Paul/Bick, Eckhard/Haltrup, Dorte 2005: Named Entity Recognition for the Mainland Scandinavian Languages. I: *Literary and Linguistic Computing 20(1)*, 91–102.

Nordgård, Torbjørn 1996: NorKompLeks: Some Linguistic Specifications and Applications. *ALLC-ACH '96*. Bergen, 214–216.

Nøklestad, Anders 2009: *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection.* PhD thesis, University of Oslo. Oslo: Acta Humaniora.

Nøklestad, Anders/Søfteland, Åshild 2007: Tagging a Norwegian Speech Corpus. I: Joakim Nivre/Heiki-Jaan Kaalep/Kadri Muischnek/Mare Koit (red.): *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, 245–248.

Oepen, Stephan/Velldal, Erik/Lønning, Jan Tore/Meurer, Paul/Rosén, Victoria/Flickinger, Dan 2007: Towards Hybrid Quality-Oriented Machine Translation. On Linguistics and Probabilities in MT. I: *The 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, 144–153.

Søfteland, Åshild/Nøklestad, Anders 2008: Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. I: Janne Bondi Johannessen/Kristin Hagen (red.): *Språk i Oslo. Ny forskning omkring talespråk*. Oslo: Novus forlag, 226–234.

Internetthenvisninger

Apertium = http://www.apertium.org/

Dokumentasjonsprosjektet = http://www.dokpro.uio.no/

EDD = http://www.edd.uio.no/index.html

Engh, Jan 1994. IBMMORF: IBM Norges leksikon og morfologi for moderne norsk. Dokumentasjonsprosjektet, Universitetet i Oslo, http://www.uio.no/~janengh/IBMmorf.htm.

Lingsoft = http://www.lingsoft.fi/

LOGON = http://www.emmtee.net/index.php?page=1&lang=no

Norsk ordbank = http://www.edd.uio.no/prosjekt/ordbanken/

Taggerprosjektet – Oslo-Bergen-taggeren = http://www.hf.uio.no/tekstlab/tagger.html

Tekstlaboratoriet = http://www.hf.uio.no/tekstlab/index.html

NorKompLeks = http://www.hf.ntnu.no/hf/prosjekter/spraktek/prosjekter/nkl

VISL = http://visl.sdu.dk/

Kristin Hagen
språkingeniør
Tekstlaboratoriet
Institutt for lingvistiske og nordiske studier
Universitetet i Oslo
Postboks 1102 Blindern
NO-0317 Oslo
kristin.hagen@iln.uio.no

Anders Nøklestad
språkingeniør, ph.d.
Tekstlaboratoriet
Institutt for lingvistiske og nordiske studier
Universitetet i Oslo
Postboks 1102 Blindern
NO-0317 Oslo
anders.noklestad@iln.uio.no

Halvautomatisk udvælgelse af lemmakandidater til en nyordsordbog

Jakob Halskov

One of the key tasks of the Danish Language Council (Dansk Sprognævn) is to monitor, record and document linguistic changes in the Danish language. To assist in this task, a prototype neologism detector called the *Ordtrawler* (Word trawler) is being developed. The system processes large amounts of text and extracts candidate neologisms using a combination of simple filters, collocational statistics and so-called neology markers. A thorough evaluation of the *Ordtrawler* indicates that neology markers are good for optimizing system precision, but at the expense of recall. Certain types of noise such as technical terms and semantically transparent compounds remain to be tackled, but diachronic frequency profiling may help.

1. Indledning

Denne artikel beskriver arbejdet med Dansk Sprognævns "Ordtrawler", en softwareprototype som har til formål automatisk at finde nyordskandidater i de meget store danske tekstkorpusser som i dag er tilgængelige på nettet. Der er foretaget en formel evaluering af dette automatiske excerperingssystem (se Halskov og Jarvad 2010), men i denne artikel er der fokus på de mere kvalitative aspekter af programmellet. Det beskrives således hvilke typer nydannelser systemet i øjeblikket kan identificere (afsnit 2), hvordan de identificeres (afsnit 3), og hvilke typer nydannelser det er mere vanskeligt at håndtere maskinelt. Endelig skitseres det hvordan systemet kan videreudvikles til at håndtere mere "avancerede" typer nydannelser ved eksempelvis at trække på nye sprogteknologiske resurser som DanNet-ontologien (afsnit 5).

1.1. Hvad er et nyt ord?

Sproglige nydannelser kan forekomme på alle lingvistiske niveauer, lige fra det semantiske niveau over de fraseologiske og syntaktiske niveauer til det ortografiske niveau. I denne sammenhæng er det primært det (leksikalsk) semantiske niveau som interesserer os, og på dette niveau kan sproglige nydannelser inddeles i de følgende tre kategorier:

1. Nye ord som refererer til nyt indhold (fx *app* om software til moderne mobiltelefoner)
2. Eksisterende ord som får nyt indhold (fx *skyde* i betydningen '(op)tage' om film)
3. Eksisterende ord som erstatter eksisterende indhold (fx *omkring* for *om*)

Kategori 1 og 2 kaldes også henholdsvis neoformativer og neosemantismer (Fjeld og Nygaard 2010). Kategori 2 og 3 kan samordnes under betegnelsen "ny brug", og denne brede betegnelse kan også omfatte ny valens eller ændrede selektionsrestriktioner og dermed involvere det syntaktiske niveau. Kategori 1 kan naturligvis også omfatte nye flerordsudtryk eller neofrasemer (Fjeld og Nygaard 2010).

Nye ord og ny brug af gamle ord er nyheder i forhold til det inventar af ord og betydninger som findes i forvejen. Mængden af opslagsord og betydninger i gængse ordbøger[1] udgør naturligvis kun en delmængde af det samlede danske almensproglige ordforråd. "Ny" bruges derfor i denne kontekst i betydningen ny for leksikografen og det leksikografiske miljø.

1 Fx Den Danske Ordbog, Retskrivningsordbogen, Nudansk Ordbog, og ikke mindst Nye ord i dansk på nettet fra 1955 til i dag (www.nyeordidansk.dk) og Dansk Sprognævns Samling.

1.2. Et spørgsmål om prægnans

Nydannelser som ikke antages at ville blive etablerede i ordforrådet, har ingen varig effekt på sproget og er dermed ikke interessante som lemmaemner i forhold til en nyordsordbog. Selvom det er vanskeligt at spå om nydannelsers fremtid, så er der kategorier af nydannelser som det, for den menneskelige excerpist, er relativt let at afvise som lemmakandidater, nemlig banale sammensætninger, lejlighedsdannelser og, i et vist omfang, "kometord" (Jarvad 1995:173).

Banale sammensætninger er fx *klimakonference*, *værdipapirsammensætning* og *risikoappetit*. Orddannelsen følger de grammatiske regler, og resultatet er semantisk transparente sammensatte ord som er helt uproblematiske at forstå hvis man kender førsteled og sidsteled. Banale sammensætninger kan ofte være relevante at indlemme i "almindelige" mono- og bilingvale ordbøger, men da de ikke har nogen nyhedsværdi, bør de ikke indlemmes i en nyordsordbog. Indlemmede man dem, ville nyordsordbogens lemmainventar hurtigt antage astronomiske dimensioner (jf. figur 1).

Lejlighedsdannelser er ofte knap så gennemskuelige, idet deres semantiske indhold er meget afhængigt af den kontekst hvori de er ytret. Til gengæld er de, for redaktøren af en nyordsordbog, lige så irrelevante som de banale sammensætninger på grund af deres forbigående karakter. Jarvad (1995) har følgende betragtninger om lejlighedsdannelser:

> Vi kan lave ord som har et mere tilfældigt præg, fx *tefest* (jf. *vinfest*), *tesøster* (jf. *kaffesøster*), *tetår* (jf. *kaffetår*), og *tetelt* (jf. *øltelt*). [...] Disse ord kaldes øjebliksdannelser eller lejlighedsdannelser, og nogle kalder dem individualdannelser. Om disse ords forståelighed er der det at bemærke at de måske nok forstås, men det bagvedliggende ord som fx *kaffesøster*, *øltelt* med disse ords bibetydninger gør forståeligheden større. (Jarvad 1995:173)

En helt særlig udfordring udgøres af ord som man umiddelbart ikke ville spå nogen særlig levetid, men som pludselig går hen og bliver meget udbredte i sprogsamfundet i en kortere periode hvorpå de så forsvinder igen. De ord som forsvinder igen, kan retrospektivt kaldes "kometord". Nogle nyere eksempler på kategorien er *burkaudvalg, klimakaravane* og *mælkeskandale*. Kometord får per definition aldrig en central position i ordforrådet, for hvis de gjorde, ville deres frekvensprofil ikke være kometagtig, altså pludseligt og kraftigt stigende fra nul og derpå hurtigt aftagende til nul igen. Alligevel kan nyordsleksikografen godt komme til at indlemme kometord i sin nyordsordbog eller udelade prægnante lemmakandidater grundet en fejlagtig antagelse om lemmaets kometstatus. Det skyldes naturligvis at det at verificere en kometagtig frekvensprofil kan fordre en tidshorisont som skal måles i år snarere end måneder.

Når lemmakandidater skal selekteres til en nyordsordbog, er den afgørende kvalitetsparameter deres anslåede prægnans, altså med hvilken sandsynlighed de kan blive varige tilskud. Som sagt, og som understreget i Fjeld og Nygaard (2010), så kræver det derfor en vis udbredelse over en vis tid at skelne prægnante nydannelser fra kometord.

> Alle nyord kan oppfattes som potensielle okkasjonalismer. [...] Bare om de bliver brukt av mange og over tid, regnes de som nyord, som seinere igjen kan bli allmennord. (Fjeld og Nygaard 2010)

Da nyordsordbøger i dag også udgives online, så er det imidlertid muligt løbende at revidere ordbogens lemmainventar og retrospektivt markere eventuelle kometord som sådanne eller helt fjerne dem. Ordtrawleren fokuserer således i første omgang på at eliminere andre former for støj, herunder de banale sammensætninger, og lader håndteringen af kometordene indgå i det fremtidige udviklingsarbejde (se afsnit 5).

2. Hvordan excerperer en maskine?

De tekstkilder hvorfra der ved Dansk Sprognævn manuelt er blevet excerperet nye ord siden 1955, omfatter aviser, ugeblade, tidsskrifter, bøger og mange andre typer tekster som repræsenterer forskellige genrer. Den halvautomatiske excerpering, som foretages af Ordtrawleren, har indtil videre taget udgangspunkt i et større antal dagblade der er tilgængelige i elektronisk form via Danmarks største artikeldatabase, InfoMedia. InfoMedia indeholder p.t. ca. 20,6 millioner avisartikler. Der er dog planer om at udvide den halvautomatiske excerpering til også at omfatte en række nyere elektroniske medier, såsom chat, blogs, Facebook og lignende.

Mens den menneskelige excerpist kan trække på sin ekspertviden om modersmålets orddannelse og ordforråd og fx anvende et omfattende antal referenceværker til at tilbagedatere nyordskandidater, så er den maskinelle excerpist på én gang mennesket underlegen og overlegen.

Underlegenheden skyldes primært at den automatiske natursprogsbehandling slet ikke kan konkurrere med menneskets. Mennesket lemmatiserer og normaliserer problemfrit enhver nyordskandidat uanset om kandidaten indgår i dets aktive ordforråd eller ej, men maskinen skal have detaljerede instrukser om enhver form for sproglig analyse og er kun i begrænset omfang i stand til at gætte på forhold vedrørende ord som den ikke har i sine ordbøger. Maskinen har stadig svært ved at abstrahere medmindre den er blevet eksplicit programmeret til det. Derfor vil den som udgangspunkt opfatte stavevarianter som vidt forskellige ord og fx foreslå *U.S.A.* som nyordskandidat selvom den har formen *USA* i sin liste over allerede kendte ord. På tilsvarende vis skal en maskine have detaljerede instrukser om hvordan den skal håndtere bindestreger, små/store bogstaver, citationstegn osv. I modsætning til menneskelige excerpister, for hvem det er helt naturligt at inddra-

ge et ords kontekst, så kræver det temmelig avancerede programmeringsteknikker at få maskiner til at tage hensyn til den sproglige kontekst hvori et ord optræder. Således vil ny brug af eksisterende udtryk, ny valens, nye flerordsforbindelser osv. være vanskelige at få en maskine til at identificere (mere herom i afsnit 5).

Overlegenheden består derimod primært i at maskinen ikke har problemer med fokusere. Den distraheres ikke af en succesoplevelse, er ikke særligt interesseret i nydannelser inden for bestemte faglige domæner eller på bestemte lingvistiske niveauer, og endelig kan den bearbejde enorme mængder empiri på kort tid og 100 % systematisk.

2.1. Hvilke excerperingssystemer findes der?

Der findes adskillige natursprogsbehandlingssystemer til automatisk detektion af termkandidater (potentielt fagsproglige udtryk) i et tekstkorpus, fx Termostat[2] (Drouin 2003), TermExtractor (Pantel og Lin 2001) og tidlige implementeringer som den beskrevet i Ahmad (1993). Til gengæld er der nærmest ikke publiceret nogen specifikationer for systemer som kan excerpere almensproglige eller fagsproglige nydannelser. Lad det dermed ikke være sagt at sådanne systemer ikke eksisterer, for det engelske APRIL-projekt (A knowledge-rich tool for the analysis and prediction of innovation in the lexicon)[3] er eksempelvis et ret velkendt, omend udokumenteret, et af slagsen. Desuden er der et enkelt helt nyt og nordisk eksempel på et automatisk excerperingssystem som er beskrevet i Fjeld og Nygaard (2010).

Ordtrawleren består i sin nuværende form af en håndfuld tekstbearbejdningsprocedurer (små programstumper), en stor database (til lagring af tekstmateriale, filtre og nyordskandidater), en korpusservice og en simpel brugergrænseflade til forskellige

2 http://olst.ling.umontreal.ca/~drouinp/termostat_web/

3 http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=GR/L08243/01

korpusværktøjer. De anvendte teknikker beskrives i hovedtræk i det følgende afsnit.

2.2. Hvordan excerperer Ordtrawleren?

Som nævnt er Ordtrawlerens primære tekstkilde faste leverancer af korte avisartikler fra InfoMedia. Denne samling af elektroniske avisartikler må underkastes en række automatiserede behandlinger i en bestemt rækkefølge før Ordtrawleren kan excerpere og altså selektere lemmakandidater til en nyordsordbog.

1. **Tokenisering**: Brødteksten deles op i sætninger, og sætningerne hakkes op i en sekvens af ordformer.
2. **Ordklassetagging**: Hver ordform tildeles automatisk en ordklasse (her anvendes de forenklede Parole-tags (Keson 1998)).
3. **Lemmatisering**: Hver ordform tildeles automatisk et lemma.
4. **Indeksering**: De enkelte oplysninger om hver ordform, dvs. formen, ordklassen og lemmaet lagres og indekseres i en database så man hurtigt kan gennemsøge store mængder tekst for bestemte mønstre (her anvendes *Corpus Workbench*-formatet[4]).
5. **Filtrering/sortering**: Inventaret af samtlige forskellige ordformer filtreres og/eller sorteres ved hjælp af ord- og frekvenslister over allerede kendte ord.

Den automatiske tokenisering, ordklassetagging og lemmatisering er naturligvis ikke fejlfri. Den anvendte tagger er beskrevet i Hansen (2000) hvor den vurderes at have en træfrate på 96,5 %, men "dog bliver kun ca. 80 % af alle ukendte ord gættet" (Hansen 2000:7). Ord som ikke er indeholdt i taggerens ordbog, volder

4 http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

altså særligt store problemer, og det er i dette tilfælde netop en delmængde af disse ord vi er ude efter. Problemet gælder i særlig grad den lemmatiseringsteknik som i øjeblikket anvendes af Ordtrawleren. Der anvendes nemlig en udfoldet version af Retskrivningsordbogen 2001 hvor samtlige bøjningsformer af hver homograf i alle ordbogsartikler genereres automatisk. Ordformer som ikke kan henføres til en homograf i dette værk, lemmatiseres dermed ikke.

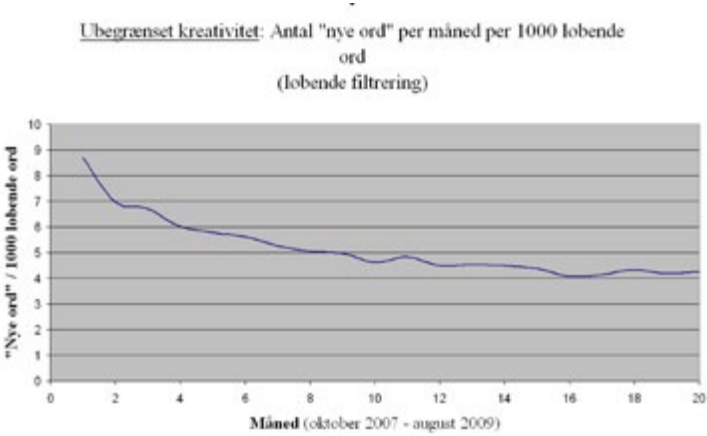
Ordform	*Ordklasse*	*Lemma*
At	UKONJ	at
forbyde	V_INF	forbyde
salg	N	salg
af	PRAEP	af
tobak	N	tobak
er	V_PRES	være
ikke	ADV	ikke
en	PRON_UBST	en
måde	N	måde
at	UKONJ	at
forlænge	V_INF	forlænge
danskernes	N_GEN	dansker
levetid	N	levetid
med	PRAEP	med
.	TEGN	.

Tabel 1: En bearbejdet sætning i det særlige Corpus Workbench-format

Tabel 1 indeholder et eksempel på hvordan en sætning ser ud efter ovenstående automatiserede behandlinger. Når alt tekstmaterialet foreligger i dette format, trækkes hele ordforrådet ud (dvs. alle *forskellige* ordformer, også kaldet "types") og sammenlignes form for form med det ordforråd systemet kender i forvejen. I afsnit 3 beskrives de teknikker systemet anvender til lemmakandidatudvælgelsen, og i afsnit 3.1 beskrives det hvilke eksisterende ordbøger og referenceværker der udgør det kendte ordforråd i denne sammenhæng.

2.3. Ordtrawlerens empiri: Ubegrænset sproglig kreativitet

Dette afsnit illustrerer den grænseløse sproglige kreativitet og produktivitet som Ordtrawleren må forsøge at navigere i. Figur 1 nedenfor viser hvordan Ordtrawleren måned for måned registrerer tusinder af ukendte ordformer i de store mængder nyhedsartikler fra InfoMedia. Det gør den også selvom man fjerner ca. 1,2 millioner allerede kendte ordformer fra et antal ordbøger og referencekorpusser og ser bort fra ikke-ord (fx e-mail-adresser) og proprier og tilmed anvender filtrering af alle hidtil observerede ordformer.



Figur 1: Ubegrænset sproglig kreativitet

Selv efter 20 måneder observerer systemet stadig mellem fire og fem ukendte ordformer per 1000 løbende ord. En enkelt måneds data fra InfoMedia (ca. 7 mio. løbende ord) bidrager således med ikke mindre end ca. 30.000 "nye ord". Citationstegnene tilkendegiver at en menneskelig excerpist naturligvis aldrig vil betragte mere end en brøkdel af disse tekststrenge som sproglige nydannelser der kan indgå i en nyordsordbog, men for en maskine er det anderledes vanskeligt at træffe sådanne afgørelser. Tallet kan virke over-

raskende, for der er relativt sjældent tale om stave- eller slåfejl i redigeret nyhedstekst, men læseren kan blot tænke på hvor mange sammensatte ord sprogbrugeren kan danne på basis af et enkelt mønster som TAL-TAL-[sejr|nederlag|...] (fx *32-14-sejren*).

Heldigvis kan man anvende et antal forskellige teknikker til enten at reducere antallet af nyordskandidater betragteligt eller til at sortere i dem på en sådan facon at de bedste kandidater kommer til at rangere højere end de dårligste. Dette er emnet for det følgende afsnit.

3. Ordtrawlerens sprogteknologiske værktøjskasse

I dette afsnit beskrives de tre teknikker som i øjeblikket anvendes af Ordtrawleren til detektion af nyordskandidater fra tekstkorpusser. Det drejer sig om henholdsvis primitiv filtrering, kollokationsstatistik og "nyhedsmarkeringer" (Jarvad 1995:23). I afsnit 4 beskrives det ganske kort hvordan disse tre teknikker med fordel kan kombineres, og hvilke resultater det har givet i en konkret formel evaluering af Ordtrawleren.

3.1. Primitiv filtrering

Selvom det hverken er tilstrækkeligt eller ideelt at fjerne et større antal allerede kendte ordformer fra analysekorpusset, så er det en helt rudimentær teknik som det i det mindste er værd at bruge lidt plads på at diskutere. Fordelen ved primitiv filtrering er at det er en meget simpel teknik som er nem at implementere. Der er dog to store ulemper ved teknikken. Dels frafiltreres der slet ikke nok uprægnante nydannelser og dels elimineres der samtidig en del prægnante nydannelser, eksempelvis ny brug af eksisterende sproglige udtryk.

Tabel 2 viser hvilke primitive filtre der anvendes i den nuvæ-

rende version af Ordtrawleren. For Den Danske Ordbog og Dansk Sprognævns Ordsamling[5] er det kun opslagsordene i deres grundform som har været tilgængelige for Ordtrawleren. Antallet af samtlige bøjningsformer der kan dannes på basis af disse grundformer, er således væsentligt højere og ville bringe den samlede størrelse af de primitive filtre langt over 1,2 mio. ordformer. For Korpus 90 og Korpus 2000 er det omvendte tilfældet. Her er antallet af lemmaer ukendt.

Nr.	Filter	Antal lemmaer	Antal ordformer
1	Retskrivningsordbogen 2001	64.038	399.062
2	I Den Danske Ordbog, men ikke i 1	34.960	34.960
3	I Ordsamlingen (september 2008), men ikke i 1-2	221.679	221.679
4	I Korpus 90, men ikke i 1-3	?	124.585
5	I Korpus 2000, men ikke i 1-4	?	436.004
I alt		?	**1.216.290**

Tabel 2: Primitiv filtrering af kendt ordmateriale

3.2. Kollokationsstatistik

En mere sofistikeret tilgang til halvautomatisk nyordsdetektion er at sortere analysekorpussets ordforråd snarere end at filtrere noget fra. Her er kollokationsstatistikken (jf. Dunning 1993) et oplagt valg, idet den med en mindre justering kan måle i hvilken grad en given ordform er særligt over- eller underrepræsenteret i et analysekorpus kontra et referencekorpus og dermed kan siges at være særligt karakteristisk for førstnævnte korpus. Selvom man normalt anvender et såkaldt (firecellet) "contingency table" (Pearson

5 Den elektroniske del af Dansk Sprognævns Ordsamling indeholder p.t. godt 273.000 opslagsord med knap 320.000 sprogbrugseksempler med udførlige kildeangivelser (belæg).

1904) til at sammenholde to ords respektive forekomster med deres samforekomst i ét korpus, kan man anvende samme firecellede tabel til at sammenholde ét bestemt ords respektive forekomst i to forskellige korpusser. De to måder at anvende "contingency tables" på er skitseret i tabel 3.

Baseret på frekvenstallene i de fire celler (kaldet O_{11} til O_{22} i tabel 3) kan man beregne et associationsmål som altså enten udtrykker tiltrækningskraften mellem to ord, fx *rejse* og *penge*, i et givet analysekorpus (jf. fodnote 8) eller mellem ét ord, fx *klimakaravane*, og to forskellige korpusser (jf. fodnote 7 og 8). Eksemplerne i tabel 3 illustrerer at tiltrækningskraften mellem *rejse* og *penge* er stærkere end mellem *rejse* og *over,* og *rejse penge* er dermed en stærkere kollokation end *rejse over*. På samme facon er nyordskandidaten *klimakaravane* mere karakteristisk for analysekorpusset, hvor den forekommer 3 gange, end for referencekorpusset, hvor den har en nulforekomst. Omvendt er *politiker* marginalt mere karakteristisk for referencekorpusset end for analysekorpusset, idet tiltrækningskraften til sidstnævnte er negativ. Tallet ligger imidlertid så tæt på 0 at ordet må siges at være lige karakteristisk for analysekorpusset (2008-2009) som for referencekorpusset (2004-2005).

En række forskellige statistiske mål kan anvendes til at beregne disse associationsstyrker, og forskellene mellem dem beror primært på hvilken vigtighed de tillægger sjældne begivenheder, altså lavfrekvensforekomster. To populære mål er henholdsvis "log-odds ratio"[6] og "log-likelihood ratio". Ordtrawleren anvender i sin nuværende konfiguration primært førstnævnte associationsmål, da dette mål tillægger lavfrekvensforekomster stor vigtighed (jf. Evert 2004). Det engelske APRIL-projekt viste nemlig at der i et vilkårligt korpus typisk vil være mange sproglige nydannelser blandt de mest lavfrekvente ord (Renouf 2002), og dermed passer "log-odds ratio" bedre end "log-likelihood ratio" til formålet.

6 log-odds-ratio = $\log((O_{11}*O_{22})/(O_{12}*O_{21}))$

	ord_2=X	ord_2≠X		ord=X	ord≠X
ord_1=Y	O_{11}	O_{12}	analysekorpus[7]	O_{11}	O_{12}
ord_1≠Y	O_{21}	O_{22}	referencekorpus[8]	O_{21}	O_{22}
Eksempel 1: "rejse penge"			Eksempel 1: "klimakaravane"		
	ord_2= penge	ord_2≠ penge		ord=klima-karavane	ord≠klima-karavane
ord_1=rejse	41	3864-41	analysekorpus	3	68373866-3
ord_1≠rejse	13849-41	39616147-3864-13849	referencekorpus	0	39616147-0
log-odds("rejse penge"): log((41*39598434)/(3823*13808)) **≈ 1,49**			log-odds("klimakaravane"): log((3.5*39616147.5)/(68373866.5*0.5)) **≈ 0.61**		
Eksempel 2: "rejse over"			Eksempel 2: "politiker"		
log-odds("rejse over"): log((13*39531012)/(3851*81249)) **≈ 0.22**			log-odds("politiker"): log(1997.5*39614853.5)/ (68371869.5*1294.5)) **≈ -0.05**		

Tabel 3: Kollokationsstatistik med "contingency tables"

3.3. Potentielle nyhedsmarkeringer

Potentielle nyhedsmarkeringer er eksplicitte sproglige eller typografiske virkemidler som ofte, men ikke altid, ledsager sproglige nydannelser i de første faser af deres liv, altså inden de for alvor måtte blive etablerede i sproget. Det er næppe muligt at give en udtømmende liste over alle potentielle nyhedsmarkeringer, men fænomenet er blandt andet omtalt i *Nye ord – hvorfor og hvordan?* (Jarvad 1995):

> En almindelig udbredelsesmåde for et nyt ord er at det dukker op i en avis, og tingen eller fænomenet bliver omtalt,

7 Dette korpus består af nyhedsartikler fra InfoMedia i perioden oktober 2008 til oktober 2009, i alt ca. 68 mio. løbende ord.

8 En større samling nyhedsartikler fra InfoMedia i perioden december 2004 til oktober 2005. I alt ca. 39,6 mio. løbende ord.

> det bliver forklaret og måske sat i gåseøjne [...] Andre nyhedssignaler er løftede kommaer omkring ordet, ordet kursiveres eller særmærkes grafisk på anden måde [...] (Jarvad 1995:23)

Set fra et sprogteknologisk perspektiv er de tre primære kvalitetskriterier for potentielle nyhedsmarkeringer henholdsvis deres individuelle frekvens og to mål som stammer fra videnskabsgrenen "Information Retrieval", nemlig "precision" (træfrate) og "recall" (genkaldelsesrate). Altså i hvor høj grad identificerer den potentielle nyhedsmarkering sproglige nydannelser og ikke andet, og i hvor høj grad kan markeringen bruges til at finde frem til alle de sproglige nydannelser der måtte være i et givet analysekorpus.

Citationstegn og kursiv er eksempler på potentielle nyhedsmarkeringer med en antaget høj genkaldelsesrate, for de kan nemlig ledsage stort set enhver type sproglige nydannelser (fx "Italien sender nu 'grundløse asylansøgere' tilbage." (Politiken 25.1.2009)). En anden prominent nyhedsmarkering er attributtet "såkaldt(e)" (fx "Meget af denne vold er såkaldt opdragelsesvold." (Politiken 2.11.2009)), men netop denne markering har den ulempe at den ikke kan ledsage VP'er og altså ikke vil kunne anvendes til at identificere nye verber.

Det er vigtigt at understrege at både de typografiske markeringer og de fleste sproglige markeringer, i særdeleshed "såkaldt(e)", naturligvis kan anvendes til at markere en række andre metasproglige forhold i teksten. Eksempelvis en ironisk, politisk eller videnskabssmæssig distance til det skrevne. I disse tilfælde vil markeringen ikke nødvendigvis pege på nydannelser, men ofte på allerede kendt ordmateriale i henholdsvis den almensproglige og den fagsproglige sfære.

Der er en række sproglige markeringer som givetvis vil have en meget høj træfrate (fx "som det hedder på nydansk"), men som grundet lav forekomst i korpus, måske er knap så anvendelige i praksis, jf. tabel 4. Omvendt er der potentielle markeringer der har

en meget høj frekvens, men som ikke i sig selv er tilstrækkeligt præcise, fx bindestreger (se tabel 4). Heidemann (2009) har undersøgt brugen af bindestreg i moderne importord og blandt andet fundet at det primært er svagt leksikaliserede ord og ord der har beholdt deres engelske udtale og skrivemåde, som særskrives. Da brugen af bindestreg er en hybrid mellem den sær- og sammenskrevne form, og da nydannelser som udgangspunkt er svagt leksikaliserede, så giver det god mening forsøgsvis at anvende bindestreg som potentiel nyhedsmarkering. Dette træk kan naturligvis ikke stå alene, men sammen med andre træk kan bindestreger signalere at der er tale om en nydannelse.

Markering	**Frekvens**	**Fremfinder:**	**Antaget effektivitet**[9]
bindestreger[10]	7366 ppm[11]	mest NP'er	Højeste frekvens, moderat genkaldelsesrate, laveste træfrate
citationstegn I (")	832 ppm	alt	Høj frekvens, højeste genkaldelse, lav træfrate
citationstegn II (')	427 ppm	alt	ditto.s.
såkaldt(e)	159 ppm	NP'er	Moderat frekvens, genkaldelse og træfrate
som den|det|de hedder	8 ppm	NP'er, VP'er	Lav frekvens, høj genkaldelse, højere træfrate
som den|det|de kaldes	0,8 ppm	NP'er, VP'er	Meget lav frekvens, høj genkaldelse, høj træfrate

Tabel 4: Potentielle nyhedsmarkeringer og antaget effektivitet

9 Den antagede effektivitet bygger på intuition og erfaring og bør naturligvis undersøges empirisk.

10 Tallet angiver antallet af ord med præcis én bindestreg.

11 Frekvenserne i tabel 4 er baseret på det korpus som er beskrevet i fodnote 6. Ppm står for "parts per million" og angiver i denne kontekst antal forekomster per million løbende ord i korpus.

Som det beskrives i næste afsnit, der omhandler en formel evaluering af Ordtrawlerens evne til at identificere sproglige nydannelser, anvendes der i øjeblikket kun den potentielle nyhedsmarkering "såkaldt(e)", markeret med gråt i tabel 4, da denne antages at repræsentere den bedste balance mellem frekvens, genkaldelsesrate og træfrate. Der er imidlertid planer om at anvende en kombination af potentielle nyhedsmarkeringer, inklusive bindestregerne, og samtidig formelt evaluere hver enkelt markerings træf- og genkaldelsesrate.

4. Formel evaluering af Ordtrawleren

I Halskov og Jarvad (2010) foretages en grundig formel evaluering af Ordtrawlerens evne til at identificere sproglige nydannelser der efter en manuel evaluering, kan optages som lemmaer i en nyordsordbog. Detaljerne i denne evaluering skal ikke gengives her, men resultatet beskrives i dette afsnit i hovedtræk.

Evalueringen omfatter to dele. Først evalueres systemets genkaldelsesrate og træfrate på et mindre analysekorpus på 75.000 løbende ord hvori samtlige sproglige nydannelser var blevet manuelt identificeret af to excerpister, og derpå evalueres træfraten på et stort analysekorpus (96,7 mio. løbende ord).

Den førstnævnte evaluering viste at den primitive filtrering reducerede de ca. 14.000 ordformer i analysekorpusset til 589 nyordskandidater hvoraf 21 % optrådte i guldstandarden, dvs. mængden af nydannelser som de menneskelige excerpister på forhånd havde identificeret i korpusset. Genkaldelsesraten var imidlertid kun 60 % hvilket illustrerer at primitiv filtrering ikke er en optimal teknik, da mange genuine nydannelser elimineres. Evalueringen viste til gengæld at kollokationsstatistikken er en bedre teknik, idet man med denne opnår en tilsvarende træfrate, men uden at filtrere nogen kandidater fra. Endelig blev det påvist at

potentielle nyhedsmarkeringer giver den højeste træfrate (ca. 40 %), men en meget lav genkaldelsesrate.

I evalueringens anden del, hvor fokus var optimering af træfrate uden hensyntagen til genkaldelsesrate, anvendtes en kombination af primitiv filtrering med de potentielle nyhedsmarkeringer. Med denne kombinerede teknik kunne der udtrækkes ca. 1800 nyordskandidater fra det store analysekorpus med en træfrate på knap 40 % og en enighedsgrad mellem de to menneskelige evaluatorer på 84,4 %. Denne træfrate er sammenlignelig med de godt 30 % som rapporteres i Fjeld og Nygaard (2010).

4.1. Evaluering af støj og reduktion af samme

Halskov og Jarvad (2010) evaluerer ikke blot Ordtrawlerens formelle træfrate og genkaldelsesrate, men undersøger også mere kvalitativt hvilke typer støj systemet har svært ved at håndtere og af forskellige årsager fejlagtigt anser for genuine sproglige nydannelser. Formålet med denne "støjanalyse" er at sondere terrænet for mulige forbedringer af systemet (mere herom i afsnit 5).

De primære støjkategorier omfatter

- Bøjningsform af kendt ord
- Banal sammensætning eller lejlighedsdannelse
- Fagsprog
- Stavefejl

De bøjede former af allerede kendte ord er den støjtype som volder Ordtrawleren de største problemer. Årsagen er at ukendte ordformer (altså ordformer som ikke kan henføres til noget lemma i Retskrivningsordbogen 2001) ikke lemmatiseres, og da opslagsordene i eksempelvis Dansk Sprognævns Ordsamling, naturligt nok, står opført i grundformen (se tabel 2), så elimineres de bøjede former

ikke. Den oplagte forbedringsmulighed er naturligvis at anvende et lemmatiseringsprogram som via grammatiske regler kan gætte på grundformen af ord der ikke optræder i dets interne ordbog.

Banale sammensætninger og lejlighedsdannelser er meget svære for Ordtrawleren at skelne fra de nye, blivende ord. Korpuseksempler på førstnævnte er *forskningskvalitet* og *fodboldekspert*, mens *kommunedans* og *pizzabande* er eksempler på de mere uigennemskuelige lejlighedsdannelser. At fjerne disse typer støj maskinelt vil være meget svært. Som beskrevet i afsnit 5 er det eneste håb sandsynligvis at anvende diakron frekvensanalyse.

At skelne fagsprog fra almensprog er, også for mennesker, en ikke-triviel opgave, og for maskinen er det særligt svært da både almensproglige nydannelser og etablerede fagsproglige udtryk ofte vil forekomme nogenlunde lige sjældent i et almensprogligt korpus. Der er dermed ikke noget oplagt statistisk kriterium med hvilket fagsprog kan filtreres fra uden at man også derved frasorterer en række genuine almensproglige nydannelser.

I forhold til støjkategorien stavefejl viste evalueringen at 65 % af disse vedrører brugen af bindestreg (fx *denial-of-service-angreb* og *nul-tolerancepolitik*). Der ligger således en oplagt systemforbedring i at implementere en simpel algoritme som ikke blot sammenholder den eksakte tekststreng for hver nyordskandidat med det allerede kendte ordforråd, men desuden tager samtlige bindestregsvarianter af nyordskandidaten i betragtning.

5. Videreudvikling af Ordtrawleren

I dette afsnit skitseres to oplagte teknikker som kunne tages i anvendelse for henholdsvis at få Ordtrawleren til at identificere mere "avancerede" sproglige nydannelser, såsom ændrede selektionsrestriktioner, og for at eliminere mere vanskelige støjtyper, fx banale sammensætninger og kometord.

5.1. Brug af DanNet til at identificere ændrede selektionsrestriktioner

DanNet-ontologien (Pedersen et al. 2009) er et almensprogligt leksikalsk-semantisk ordnet som er opbygget efter forbillede af Princeton WordNet (Fellbaum 1998), og som siden marts 2009 har været frit tilgængeligt som elektronisk resurse via hjemmesiden http://wordnet.dk.

DanNet består af et antal "synsets" (bundter af synonymer som tilsammen kan anses for at udpege et bestemt begreb), et antal semantiske relationer (fx "used_for" og "has_hyperonym") og endelig et antal ontologiske typer som hvert synset kan siges at repræsentere. Grundlæggende er der to hierarkier af ontologiske typer i DanNet, nemlig

1. konkrete genstande ("1st Order Entities")
2. handlinger, hændelser, egenskaber og abstrakte genstande ("2nd Order entities" og "3rd Order entities")

Ontotypen "Artifact" er en "1st Order Entity" som realiseres i sproget af eksempelvis et ord som *kosmetik*. Ontotypen "Property" er en "2nd Order Entity" (fx *glæde* eller *tørhed*), og "Time" er en "3rd Order Entity". De fleste ontotyper som har fundet konkret anvendelse i DanNet (der er omkring 190 forskellige i version 1.1 af resursen), er imidlertid sammensatte ontotyper, fx er synsettet *kniv* klassificeret som "Instrument+Artifact+Object".

I dette afsnit fokuseres der på disse ontologiske typer og deres potentiale i forhold til at identificere ændrede selektionsrestriktioner for et givet lemma i et analysekorpus kontra et (ældre) referencekorpus.

Den grundlæggende hypotese er at man ved at analysere hvilke ontologiske typer et givet lemmas konkrete syntaktiske argumenter repræsenterer på forskellige tidspunkter i et diakront korpus,

kan få et indtryk af om der sker signifikante forskydninger hen mod eller væk fra bestemte grupper af ontotyper. Er dette tilfældet, så repræsenterer ændringerne, antageligvis, ændringer i lemmaets selektionsrestriktioner.

Analysen er i første omgang meget simplistisk og indebærer simpelthen automatiske opslag i DanNet af samtlige lemmaer der optræder i position +1 i forhold til inputlemmaet (en approksimation af den prototypiske objektsposition), og efterfølgende optælling af hvilke, om nogen, ontologiske typer disse lemmaer kan grupperes under. I nærværende artikel skitseres blot et enkelt eksempel på resultatet af denne simplistiske analyse for inputlemmaet *rejse*. Som det ses i tabel 5, har *rejse* tilsyneladende ikke undergået signifikante forandringer i forhold til de selektionsrestriktioner det lægger på sit direkte objekt. Ændringerne er i hvert fald ikke synlige fra perioden 2004-2005 til perioden 2008-2009.

2004-2005			2008-2009		
Rang	**Ontotype**	**Eksempler**	**Rang**	**Ontotype**	**Eksempler**
1	3rdOrder Entity+ Quantity	krav, kapital, penge	1	3rdOrder Entity+ Quantity	krav, kapital, penge
2	Unbounded Event+ Agentive+ Communication+ Social	tiltale, spørgsmål, debat, diskussion	2	Unbounded Event+ Agentive+ Communication+ Social	tiltale, spørgsmål, debat, diskussion
...	...	...	...	...	...

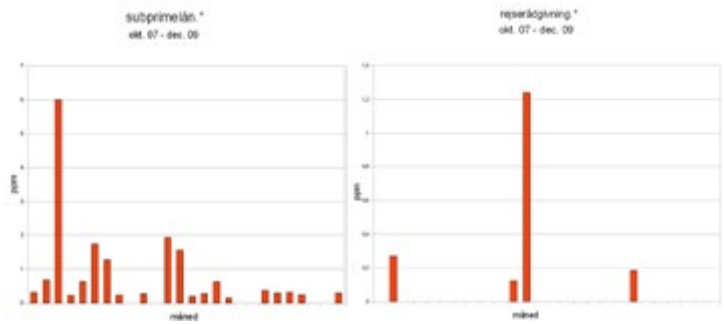
Tabel 5: Brug af DanNet til sammenligning af selektionsrestriktioner over tid: Ontotyper i prototypisk O-position for *rejse*

De to hyppigste ontologiske typer for objektargumenter i 2008-2009 er de samme som i 2004-2005, nemlig abstrakt kvantitet (kapital, penge) og uafgrænsede hændelser af social/kommunikativ karakter (spørgsmål, debat).

Der er naturligvis en række svagheder ved denne teknik. De lemmaer som ikke figurerer i DanNet, bliver eksempelvis ikke henført til nogen ontotype, og desuden er der strengt taget ikke tale om syntaktiske argumenter. Korpus er nemlig ikke er parset, og der anvendes i stedet en simpel positionel SVO-skabelon hvor inputlemmaet er V og lemmaet på position +1 i forhold til V antages at fungere syntaktisk som O. Teknikken skal derfor raffineres således at der som minimum anvendes "phrase chunking". Dermed er det i det mindste kerneleddet i den NP som følger umiddelbart efter V, der slås op i DanNet og ikke det ord som tilfældigvis står i position +1. Under alle omstændigheder virker teknikken anvendelig. Udfordringen består naturligvis i at man ikke på forhånd kan vide hvilke verber der måtte ændre selektionsrestriktioner, og det vil derfor være nødvendigt fuldautomatisk at udtrække og sammenligne data som illustreret i tabel 5 for eksempelvis de 100 eller 1000 mest frekvente danske verber. Denne videreudvikling af Ordtrawleren er på tegnebrættet.

5.2. Inddragelse af diakrone frekvensoplysninger

I dette afsnit skitseres det støjreducerende potentiale i diakrone frekvensprofiler, dvs. afbildninger af den hyppighed hvormed en given nyordskandidat forekommer i et givet korpus som repræsenterer et antal sekventielle tidsperioder (InfoMedia-tekster fra perioden oktober 2007 til december 2009). Figur 2 illustrerer hvordan to nyordskandidater, nemlig *subprimelån* og *rejserådgivning*, har meget forskellige "fingeraftryk" når deres forekomster i et diakront korpus afbildes grafisk.



Figur 2: Diakrone frekvensprofiler

For det første er den relative frekvens af *subprimelån* (0-6 ppm) markant højere end *rejserådgivning* (0-1 ppm). For det andet har *subprimelån* en kometagtig frekvens i begyndelsen af perioden (december 2007) hvorefter frekvensen aftager forholdsvis hurtigt, men dog ikke går i nul. Frekvensprofilen for *rejserådgivning* har et mere cyklisk udseende med nulforekomst i længere perioder (7-9 måneder) afløst af en vis, omend lav, forekomst i enkeltstående måneder.

Hvad kan man så udlede af disse profiler? Sund fornuft fortæller os at *rejserådgivning* er en sæsonbetonet nydannelse, og at *subprimelån* er et kometord som dog næppe forsvinder helt før den nærværende finanskrise er slut. Konklusionen må være at der skal forholdsvis lange tidshorisonter til før diakrone frekvensprofiler kan anvendes til at afsløre kometord. Horisonterne skal sandsynligvis tælles i år snarere end måneder, men selv med sådanne horisonter vil det være vanskeligt at opstille operationelle kriterier som kan danne basis for at en maskine kan afvise, respektive godkende, en nyordskandidats diakrone frekvensprofil.

Omskrives histogrammet for *subprimelån* til en funktion (efter forbillede af Fjeld og Nygaard, 2010), er resultatet tilnærmelsesvist en aftagende eksponentiel funktion, men bør man mekanisk udelukke alle nyordskandidater som følger dette mønster? Histo-

grammet for *rejserådgivning* kunne omskrives til en stykvis funktion som er henholdsvis stigende og aftagende, hvilket måske taler imod udelukkelse.

Selvom kometord kræver en vis tidshorisont at detektere, så vil mange banale sammensætninger givetvis kunne elimineres med diakrone frekvensprofiler hvor diakronien ikke nødvendigvis er omfattende. For eksempel vil sammensætningen *108-116-nederlaget* næppe forekomme i flere forskellige tidsperioder, og selv hvis den gjorde, så ville en minimumsfrekvens på eksempelvis 1 ppm i mindst én periode udelukke den (men hverken *rejserådgivning* eller *subprimelån*).

6. Konklusion og perspektiver

Artiklen illustrerer at det ingenlunde er enkelt at få en maskine til at assistere i arbejdet med at selektere lemmakandidater til en nyordsordbog. Maskinen mangler selvsagt den modersmålskompetence og omfattende omverdensviden og intuition som den menneskelige excerpist er i besiddelse af. Til gengæld kan den excerpere 100 % objektivt og bearbejde store mængder tekst langt hurtigere.

Konklusionen er dog at Ordtrawleren på sit nuværende udviklingsstadium kan udgøre en mærkbar hjælp i forhold til selektionen af de mere simple sproglige nydannelser, dvs. helt nye ord som refererer til nyt indhold. Systemet kan ikke i øjeblikket identificere flerordsforbindelser, nye valensmønstre eller nye selektionsrestriktioner, men som skitseret i afsnit 5, så kan en række sprogteknologiske teknikker og nye resurser tages i anvendelse og sandsynligvis (delvist) udbedre disse mangler.

Litteratur

Ahmad, K. 1993: Pragmatics of specialist terms: The acquisition and representation of terminology. *Machine Translation and the Lexicon, 3rd EAMT Workshop proceedings*.

Drouin, P. 2003: Term Extraction Using Non-technical Corpora as a Point of Leverage. I: *Terminology.* 9(1): 99-117.

Dunning, T. 1993: Accurate methods for the statistics of surprise and coincidence. I: *Computational Linguistics*. 19:1: 61-74.

Evert, S. 2004: *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.d.-afhandling, Stuttgart Universitet.

Fellbaum, C. (red.) 1998: *WordNet: An Electronic Lexical Database*. Boston: MIT Press.

Fjeld, R. V. og L. Nygaard 2010: Neologismer i norsk. Kartlegging av leksikalsk språkendring før og nå. I: *Nordiske studier i leksikografi.* Bind 10: 506-521. Rapport fra Konferensen om lexikografi i Norden Tammerfors 3.-5. juni 2009, red. H. Lönnroth og K. Nikula. Tammerfors 2010.

Halskov, J. og P. Jarvad 2010: Human og maskinel excerpering af neologismer. I: *Nydanske Sprogstudier (NyS)* 38, 39-68.

Hansen, D. H. 2000: *Træning og brug af Brill-taggeren på danske tekster*. Teknisk rapport fra Center for Sprogteknologi (CST). http://cst.dk/online/pos_tagger/Brill_tagger.pdf.

Heidemann, M. 2009: Om særskrivning, sammenskrivning og brugen af bindestreg i moderne importord. I: *Nyt fra Sprognævnet*. 2009/4: 28-33.

InfoMedia. Danmarks største artikeldatabase. http://www.infomedia.dk.

Jarvad, P. 1995: *Nye ord – hvorfor og hvordan?* København: Gyldendal.

Keson, B. 1998: *Vejledning til det danske morfosyntaktisk taggede PAROLE-korpus*; Teknisk rapport fra Det Danske Sprog- og Litteraturselskab (DSL). http://korpus.dsl.dk/paroledoc_dk.pdf.

Pantel, P. og D. Lin 2001: A Statistical Corpus-Based Term Extractor. I: *Lecture Notes in Computer Science.* 2056: 36-46.

Pearson, K. 1904: On the theory of contingency and its relation to association and normal correlation. *Drapers' Company Research Memoirs, Biometric Series 1*. London: Dulau & Co.

Pedersen, B.S., S. Nimb, N., J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen 2009: DanNet: the Challenge of Compiling a Wordnet for Danish by Reusing a Monolingual Dictionary. I: *Language Resources & Evaluation*. 43:3: 269-299.

Renouf, A. 2002: The Time Dimension in Modern Corpus Linguistics. I: Bernhard Kettemann & Georg Marko (red.) *Teaching and Learning by Doing Corpus Analysis. Papers from the 4th International Conference on Teaching and Learning Corpora*. 27-41.

Jakob Halskov
forsker, ph.d.
Dansk Sprognævn
H.C. Andersens Boulevard 2
DK-1553 København V
jhalskov@dsn.dk

KTHs morfologiska och lexikografiska verktyg och resurser

Viggo Kann

During the last 15 years the human language technology group at KTH has developed tools and resources that may be of interest to the lexicographical community. Several tools have been developed as part of the group's research on Swedish authoring tools: spelling error detection and correction, grammar checking, part-of-speech tagging, lemmatization, compound splitting, and an interactive learning environment called Grim. Most of the tools are open source and may be downloaded from <www.csc.kth.se/theory/humanlang>. We have also made several dictionaries available on the web: the Lexin series of dictionaries for 15 languages, the Scandinavian Dictionary, the Tvärslå dictionary collection, the Swedish Hyphenation Dictionary and the two crowdsourced resources The People's Dictionary of Synonyms and The People's English-Swedish Dictionary.

1. Språkteknologigruppen på KTH

Språkteknologigruppen är en tvärvetenskaplig forskargrupp inom avdelningarna för teoretisk datalogi och människa-datorinteraktion som hör till skolan för datavetenskap och kommunikation på KTH. Jag har lett gruppen från dess start för 15 år sedan. Vi arbetar med utbildning, forskning och utveckling i språkteknologi. Forskningen har finansierats av bland annat KTH, HSFR (Humanistisk-samhällsvetenskapliga forskningsrådet), Nutek, Vinnova, Vetenskapsrådet, Nordiska ministerrådet, Språkrådet och .SE-stiftelsen. Fem doktorander har doktorerat och disputerat helt inom gruppen. Vi arbetar i nära samarbete med språkteknologiforskarna på institutionen för data- och systemvetenskap vid Stockholms universitet och med talteknologigruppen vid KTH.

Våra huvudområden inom språkteknologi är svensk språkgranskning, informationssökning och informationsextraktion samt ordböcker. Vi vill att det vi arbetar med ska komma till användning och vara nyttigt för både allmänheten och andra forskare. Gruppens filosofi kan sammanfattas i en mening: Vi utvecklar effektiva och resurssnåla metoder för språkteknologiska system, i synnerhet för svensk text men också språkoberoende metoder, där resultaten, både algoritmer, program och resurser, görs fritt tillgängliga i möjligaste mån. Gruppens medlemmar och verksamhet inom forskning och utbildning beskrivs på <www.csc.kth.se/theory/humanlang> där också program och resurser finns för nedladdning.

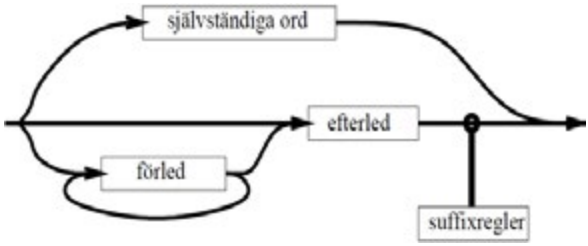
I denna artikel beskrivs och diskuteras verktyg och resurser med lexikografisk relevans som gruppen utvecklat under dessa femton år.

2. Stava – svensk stavningskontroll

Det första språkteknologiska problem vi studerade var svensk stavningskontroll. Vi konstruerade stavningskontrollprogrammet Stava, som likt de flesta sådana program är kontextlöst, det vill säga det kontrollerar varje ord separat, utan att ta hänsyn till vilka ord som står före och efter i texten. Detta gör att Stava aldrig kan upptäcka en felstavning som råkar sammanfalla med ett annat ord, till exempel om *spion* felstavas som *pion*. Eftersom vi vill bygga effektiva system som kan klara av realistiska indata har vi valt att använda SAOL (1986) som grundordlista. Ordlistan är bearbetad och lagrad så att Stava ska klara av att känna igen alla böjningsformer och sammansättningar på ett mycket effektivt sätt.

Ordlistan är uppdelad i tre delar: självständiga ord, förled och efterled.

1. Ordlistan med **självständiga ord** består av de ord som inte kan ingå i sammansättningar, till exempel *ömsom*, *eller* (1400 ord).
2. I **förledsordlistan** ingår alla förled i sammansättningar, såsom *korv*, *medie*, *packnings* (23 000 ord).
3. **Efterledsordlistan** består av alla ord som kan vara efterled i sammansättningar, till exempel *kunskap*, *stort*, *låta*, *medium*. Orden förekommer i efterledsordlistan i de former som står i SAOL. För substantiv är det till exempel grundform, bestämd form och pluralform (100 000 ord).



Figur 1: Schema för hur ord kan bildas som hopslagningar av ord

Varje ord ur ordlistan med självständiga ord och ur efterledsordlistan är tillåtna ord. Sammansatta ord kan bildas med hopslagning (konkatenering) av ett eller flera ord ur förledsordlistan och ett ord ur efterledsordlistan, se figur 1. Exempel med ordleden ovan: *korv-packnings-kunskap*. SAOL är grunden till vilka ord som hamnat i vilken ordlista, men vi har gjort vissa avvikelser. Till exempel har alla substantiv lagts i efterledsordlistan även om de inte förekommer som efterled i några sammansättningar i SAOL. Huvudsaken för Stava är att det är teoretiskt möjligt för ordet att bilda ett efterled.

Böjningsformer som inte förekommer i efterledsordlistan genereras med suffixregler, omkring 1500 stycken. Exempel på en suffixregel:

-ornas ← -a, -an, -or

Denna regel säger att om ett ord förekommer med suffixen *-a*, *-an* och *-or* så ska även suffixet *-ornas* godkänns. Exempelvis finns *docka*, *dockan* och *dockor* i efterledsordlistan, varför *dockornas* godkänns. Det finns 1800 substantiv som passar in på denna regel.

Varje ordlista lagras som ett Bloomfilter (Bloom 1970) vilket är ett extremt kompakt och effektivt lagringssätt där varje ord representeras av ett antal till synes slumpmässigt utplacerade ettor i en binär datastruktur. Med denna lagring går det mycket snabbt att avgöra om ett ord finns med i ordlistan. Ytterst sällan, med en sannolikhet som kan väljas i förväg, gör algoritmen fel och godkänner ett ord som inte finns i ordlistan. Detta sker så sällan att det inte betyder något i praktiken, men det gör att det inte går att utvinna ursprungsordlistan, något ordlisteleverantörer i allmänhet kräver. Enda sättet att skydda ursprungsordlistan helt är faktiskt att stavningsalgoritmen svarar fel ibland (Kann, Domeij, Hollman och Tillenius 2001). Annars går det nämligen att utvinna ordlistan genom att ett program testar alla möjliga bokstavskombinationer mot stavningsalgoritmen.

Att dela upp SAOL i de tre ordlistorna och att skapa suffixreglerna var ett stort arbete. Sammansättningsgränserna i SAOL var inte uppmärkta på ett entydigt sätt och på många ställen var kodningen inkonsekvent. Under arbetet fann vi över 500 fel i SAOL, som alltså hade undgått mänskliga korrekturläsare. Vi skickade felen till SAOL-redaktionen men fick tyvärr ingen respons. I senare upplagor verkar dock felen vara åtgärdade.

Stavas kodade ordlistor och källkod finns fritt tillgängliga för nedladdning (STAVA). På samma webbsida finns också en webbversion av Stava som även kan stavningskontrollera webbsidor.

3. Utvidgningar av Stava

Stava har genom årens lopp vidareutvecklats till att bli ett allt mer komplett morfologiskt verktyg. Här beskrivs de viktigaste tilläggen.

3.1. Rättstavning

Första utvidgningen som gjordes var generering av rangordnade rättelseförslag. Rättelseförslag till ett felstavat ord tas fram genom att Stava går igenom alla tänkbara ord som skulle ha kunnat ge upphov till felstavningen och kontrollerar om orden existerar. Damereau (1964) har visat att 80 procent av alla mänskliga felstavningar beror på antingen omkastning av två intilliggande bokstäver eller insättning, borttagning eller utbyte av en bokstav. Om Stava inte hittar något korrekt ord med en tillämpning av Damereaus regler så provas två tillämpningar.

De genererade förslagen rangordnas sedan i trolighetsordning med hjälp av ett felstavningsavstånd och ordfrekvenser. Felstavningsavståndet visar exempelvis att dubbelskrivning av en konsonant eller byte av en bokstav mot en intilliggande på tangentbordet ligger närmare det rättstavade ordet än andra fel. Att ordfrekvenser förbättrar rangordningen beror på att det är troligare att det är ett vanligt ord som råkat bli felstavat än ett ovanligt. På detta sätt genererar Stava rättelseförslag som är mycket bra. En undersökning har visat att 60 % av de felstavade orden i en text fick korrekt förstahandsrättelseförslag (Kann, Domeij, Hollman och Tillenius 2001). Vid en jämförelse mellan Stava, Microsoft Words rättstavningsmodul och Unixverktyget Ispell hade Stava klart bäst rättstavning (Bigert 2005).[1]

1 Korrekt rättelseförslag föreslogs i 97 % av fallen av Stava, 93 % av Ispell och 89 % av MS Word. Den korrekta rättelsen kom som för-

3.2. Ordklasstaggning och lemmatisering

Det är svårt att få tag i stora lexikon där orden är märkta med både ordklass och böjningsform. Det visade sig att vi med Stavas suffixregler kunde få ett sådant lexikon med minimalt arbete. Det enda som krävdes var att vi märkte varje suffixregel med vilken tagg den motsvarar. Exempelvis märktes exempelregeln i avsnitt 2 med *nn.utr.plu.def.gen*, vilket säger att alla 1800 ord som genereras med denna regel är substantiv i utrum, plural, bestämd form, genitiv.

På detta sätt blir Taggstava en ordklasstaggare för alla regelbundet böjda svenska ord, både enkla och sammansatta (TAGGSTAVA). Det är bara omkring 3 000 av de 100 000 orden i efterledsordlistan som inte blir taggade med denna metod, och dessa skulle det enkelt gå att ta hand om för hand. Taggstava tittar liksom Stava bara på enskilda ord utan kontext. Om man vill bestämma vilken av de tänkbara taggarna för ett ord som ordet har i sitt sammanhang så ska man använda en *disambiguerande ordklasstaggare*, såsom Granskatagger i avsnitt 4.

Med suffixreglernas hjälp kan även lemmaformen enkelt fås fram för alla ord som kan taggas, eftersom lemmaformens ändelse står först i regelns högerled. Lemmatisering kommer alltså på köpet.

3.3. Sammansättningsanalys

Med hjälp av förleds- och efterledsordlistorna kan Stava, som vi beskrivit i avsnitt 2, ta fram vilka ordled ett sammansatt ord består av. Många sammansatta ord får flera sammansättningsanalyser av Stava, och det vore värdefullt om ett system skulle kunna disambiguera bland dessa.

---

sta förslag i 88 % av fallen för Stava, 67 % för Ispell och 60 % för MS Word.

Vi har undersökt åtta metoder för disambiguering av flertydiga sammansättningar och kombinationer av dessa metoder (Sjöbergh och Kann 2006). Den bästa metoden var en statistisk kombinationsmetod som bygger på följande tre enskilda metoder:

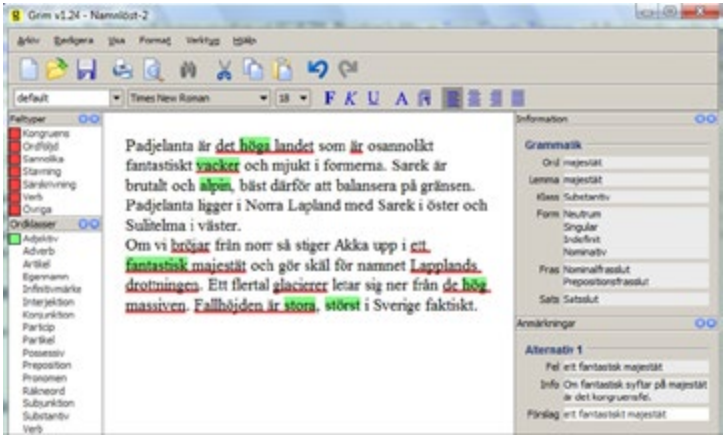
1. Välj sammansättningen med lägst antal ordled (*mun-vinklarna* bättre än *mun-vin-klarna* eftersom det bara har två ordled).
2. Välj sammansättningen med vanligast ordled (*upp-rättar* bättre än *upprätt-ar* eftersom *upp* och *rättar* är vanligare än *upprätt* och *ar*).
3. Välj sammansättningen med vanligaste kombinationen av ordledsordklasser (*upp-rättar* bättre än *upprätt-ar* eftersom preposition-verb är vanligare i sammansättningar än adverb-substantiv).

Denna kombinationsmetod delar upp 98 % av sammansatta ord korrekt. Sett över alla ord i den analyserade texten så gör sammansättningsuppdelaren bara fel på 0,1 %. Källkoden (skriven i programspråket C++) till sammansättningsuppdelaren är fritt tillgänglig.

4. Granska – svensk grammatikkontroll

Efter stavningskontrollprojektet angrep vår grupp grammatikgranskningsproblemet, något som är mycket svårt. Det finska företaget Lingsoft, som har gjort grammatikkontrollen i Microsoft Word, är den ledande producenten av svensk grammatikkontroll, men den är långt ifrån heltäckande, till exempel detekteras inte särskrivningar.

Vårt eget system Granska består till att börja med av en disambiguerande ordklasstaggare med lemmatiserare och ordböjare. Den kallas Granskatagger och finns fritt tillgänglig tillsammans med ett fritt lexikon.



Figur 2: Grim. Användaren, som just valt att alla adjektiv ska markeras, pekar på ordet *majestät* på rad 7 och får då upp analysen i fältet till höger.

Granska implementerar ett ganska avancerat grammatikregelspråk i vilket vi skrivit regler för bland annat särskrivningsdetektion. Det visade sig kunna användas också för att skriva en så kallad grund parser som analyserar och delar upp meningar i satser (Knutsson, Bigert och Kann 2003). Granska koncentrerar sig på svåra grammatiska fel och innehåller inte metoder för att hitta enklare fel såsom upprepade ord eller felaktiga skiljetecken, som till exempel Lingsofts grammatikkontroll hittar.

Vi har samlat flera av våra verktyg i ett gemensamt gränssnitt som kallas Grim (GRIM). Resultatet är en interaktiv lärmiljö som består av en ordbehandlare med stavningskontroll, regelbaserad och probabilistisk grammatikkontroll, presentation av ordklasser, sökning i lexikon med mera, se figur 2. Grim är avsett att användas i undervisning i svenska i olika stadier.

5. Nätordböcker

Gruppen har lagt upp en stort antal ordböcker på nätet. Några av dessa beskrivs här.

5.1. Lexin

Redan 1994 fick jag i uppdrag att lägga upp Skolverkets Lexin-lexikon på webben. I början av 1995 låg svensk-engelska och svensk-finska Lexin sökbara på webben, som de första svenska lexikonen på Internet. Lexins svenska ordbas består av 30 000 ord som valts för att utgöra ett lämpligt ordförråd för den som flyttar till Sverige. Den svenska ordbasen är numer översatt till 15 språk: albanska, arabiska, bosniska, engelska, finska, grekiska, kroatiska, nordkurdiska, persiska, ryska, serbiska, somaliska, spanska, sydkurdiska och turkiska. Alla dessa finns sökbara på webben (LEXIN). Stavas rättstavningsalgoritm finns inlagd i Lexin, varför felstavade sökningar på både källspråk och målspråk rättas automatiskt om de är entydiga. Om det finns flera möjliga rättelser får användaren möjlighet att välja mellan dessa. Dessutom har Lexins ordbas översatts till norska, danska och isländska och finns på webben i alla fyra länder.

Antalet uppslagningar i svenska Lexin har ökat år från år och är nu omkring 20 miljoner i månaden. Mängder av användarstatistik kan därmed samlas in. Denna statistik kan användas på flera sätt. När Skolverket skulle sätta ihop ett engelsk-svenskt lexikon tog vi fram en lista på de 40 000 vanligaste orden som användare slagit upp men som saknades i Lexin. Dessa ord översattes och blev därmed ett utmärkt komplement till Lexin, speciellt avpassat till det som användare brukar vilja slå upp. Statistiken kan också, tillsammans med användarenkäter, nyttjas för att se hur ordböcker används (Hult 2008).

Ägarskapet till Lexin övergick från Skolverket först till Myndigheten för skolutveckling och därefter till Språkrådet.

5.2. Skandinavisk ordbok

Skandinavisk ordbok utvecklades 1994 av Nordiska språksekretariatet inom Nordiska ministerrådet och består av 10 000 lemman som skiljer mellan de skandinaviska språken svenska, danska och norska. Vi gjorde en webbversion av detta lexikon för över tio år sedan (SKANDORD). Ordboken innehåller bara orden, inga definitioner eller annan information.

5.3. Tvärslå

Det nordiska samarbetsprojektet Nordisk nätordbok finansierades av Nordiska ministerrådet 2005–2007. I projektet skapades ett bibliotek av flerspråkiga nätordböcker mellan minst två av de nordiska språken och engelska. Dels skapades några nya flerspråkiga ordböcker med automatiska eller halvautomatiska metoder, dels samlades existerande ordböcker ihop: svensk-engelska Lexin, engelsk-svenska Lexin, svensk-finska Lexin och Skandinavisk ordbok var med bland dessa ordböcker. Ordböckerna har lagrats i ett för projektet speciellt framtaget XML-format, så att de skulle kunna användas på ett enkelt och enhetligt sätt i bland annat forskningssammanhang. Inom projektet tog vi också fram söktjänsten Tvärslå, i vilken man kan slå upp i alla ingående ordböcker samtidigt. Antingen kan man slå upp på alla språk samtidigt eller på ett specificerat språk (TVÄRSLÅ).

5.4. Avstavningslexikon

Boken Svenskt avstavningslexikon (Klingspor 1985) är utgången från förlaget, och Språkrådet som har rättigheterna till materialet

har beslutat att göra lexikonet tillgängligt på webben. Det handlar om 33 000 lemman och ungefär 100 000 avstavade ordformer. Materialet har skannats in, vilket gör att många fel införs. Med ekonomiskt stöd av Ebba Danelius stiftelse har vi automatgranskat och rättat det inskannade materialet från boken och tagit fram en rudimentär söktjänst som kommer att utvecklas vidare (AVSTAV).

6. Utveckling av fria nätordböcker

Ett problem inom språkteknologisk forskning och utveckling är att det inte finns så många tillgängliga språkteknologiska resurser, såsom ordböcker, korpusar och trädbanker. En anledning är att det kräver mycket manuellt arbete att sätta ihop en stor resurs, och att den som gjort det vill ha den för sig själv eller ta betalt för resursen. Detta gäller ofta även resurser som tagits fram i projekt finansierade med statliga medel. Även för resurser som är tillgängliga gratis gäller ofta att det finns begränsningar av hur de får användas och ändras.

Inom programutvecklingsvärlden finns det en stark rörelse för fri programvara. Den härstammar ur hackerkulturen, men allt fler företag och individer har insett att fri programvara är något som alla har glädje av och som kan snabba på utvecklingen av nya och existerande program. Kända exempel på fri programvara är GNU-projektets verktyg och operativsystemet Linux.

På samma sätt är fria språkresurser något som skulle gynna utvecklingen av nya språkteknologiska system. En fri språkresurs skulle kunna definieras på följande sätt, med en lätt modifiering av GNU-projektets grundare Richard Stallmans definition av fri programvara (Stallman 1996):

1. Frihet att använda resursen i en tillämpning.
2. Frihet att studera resursen och modifiera den.

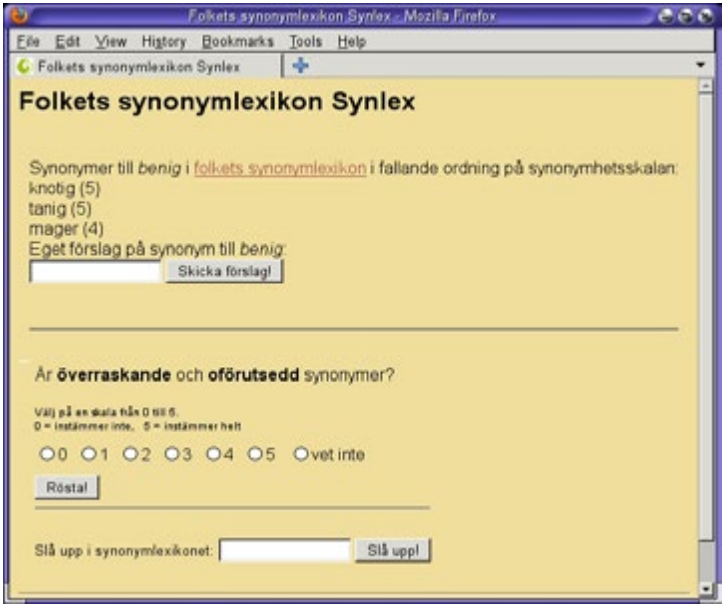
3. Frihet att ta en kopia av resursen.
4. Frihet att förbättra resursen och låta andra få kopior av den förbättrade resursen.

Är en ordbok som man kan slå upp i på nätet en fri resurs? Nej, enligt ovanstående definition är det inte nödvändigtvis så. För att den ska vara en fri resurs måste det också gå att ladda ner den, studera den och ändra den. Hur kan man då lösa upphovsrättsproblemet och skapa en fri resurs utan att några få personer behöver lägga ner enorma mängder ideellt arbete? Jo, det finns språkteknologiska metoder som automatiskt kan skapa stora resurser som visserligen till delar är skräp, men där mycket av innehållet är korrekt. Om det omfattande granskningsarbetet av en sådan resurs kan fördelas på många människor så behöver varje persons bidrag inte vara så stort. Detta arbetssätt att utnyttja folket till att gemensamt konstruera något kallas med ett nytt engelskt ord för *crowdsourcing*.

6.1. Folkets synonymlexikon

Vi har använt den ovan nämnda metoden för att skapa ett svenskt synonymlexikon (Kann och Rosell 2005). Vi började med att automatiskt konstruera en massa ordpar som skulle kunna vara synonymer. Detta gjorde vi genom att med hjälp av elektroniska ordböcker översätta svenska ord till engelska och sedan tillbaka till svenska igen. Därigenom hittas många äkta synonymer men också mängder av falska synonymer på grund av homonymi. Dessa ordpar rensade vi sedan med hjälp av Random indexing, en statistisk metod som grundar sig på i vilka kontexter orden förekommer; ord som förekommer i liknande kontext är förmodligen lika varandra, och ord som är lika varandra skulle kunna vara synonymer.

Därefter lät vi tiotusentals människor bidra genom att kontrollera synonymiteten hos ordpar. Till detta fick vi tillstånd att an-



Figur 3: Användargränssnitt för Folkets synonymlexikon

vända Lexins webbplats under ett par månader. Under denna tid fick alla som gjorde en uppslagning i Lexin också upp en fråga om ett ordpar, se nedre delen av figur 3 för ett exempel. Användarnas ordparsbedömningar analyserades och de par som fick goda omdömen lades in i synonymlexikonet.

På kort tid skapades härigenom ett svenskt synonymlexikon (SYNLEX). Därefter togs frågorna bort från Lexins webbplats, men en länk till Folkets synonymlexikon behölls. Den som vill slå upp i lexikonet behöver göra minst en ordparsbedömning först. Det går också bra att föreslå egna synonymer som sedan bedöms av andra användare. På detta sätt växer synonymlexikonet hela tiden i omfattning och kvalitet. I januari 2010 hade 3,6 miljoner bedömningar gjorts. Det fanns då 80 000 ordpar (som i medeltal bedömts med minst 2 på den femgradiga skalan). Totalt hade

123 000 ordpar föreslagits av användare, varav 74 000 olika. 24 000 av dem hade dittills accepterats av andra användare.

En finess med Folkets synonymlexikon jämfört med vanliga synonymlexikon är att synonymerna är graderade mellan 2 och 5, genomsnittet av alla användares bedömningar. Det är alltså varje användares egen intuitiva uppfattning om begreppet synonymitet som bestämmer hur ett ordpar graderas, se till exempel figur 3. Det skapade lexikonet bygger således på folkets egen definition av synonymitet, vilket förhoppningsvis är precis vad folket vill! Eftersom lexikonet är skapat av hela folket så är det en fri resurs som kan laddas ner i sin helhet.

Baksidan med att använda allmänheten som arbetskraft är att det förekommer individer som försöker förstöra. Alla projekt som bygger på crowdsourcing måste hantera detta på något sätt. Vi har på följande tre sätt försäkrat oss mot att missbruk av lexikonet ska påverka kvaliteten:

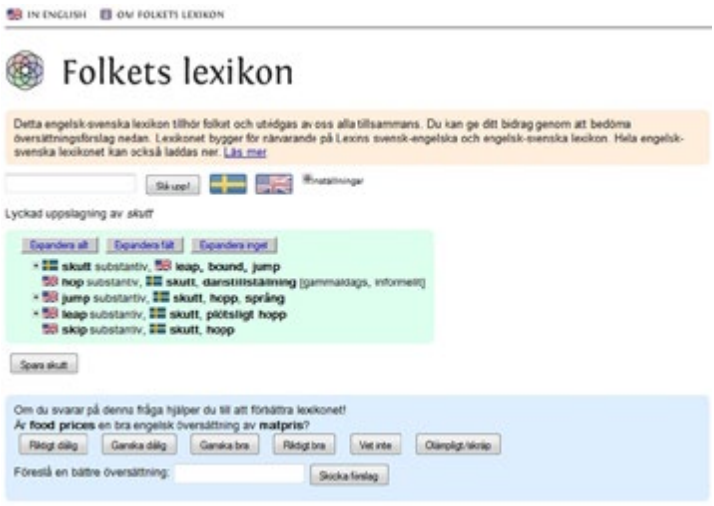
1. Många bedömningar krävs innan ett ordpar anses vara bra.
2. Vilket ordpar som ska bedömas väljs slumpmässigt från en lista som består av nästan 100 000 par.
3. Ordpar som föreslås av användare stavningskontrolleras innan de läggs till den enorma listan.

Det tycks som att dessa försäkringar är starka nog för att synonymlexikonets kvalitet ska hållas tillräckligt hög. En stickprovsundersökning där Folkets synonymlexikon jämfördes med en tryckt synonymordbok visade att alla synonymer som graderats 3 eller högre var angivna som synonymer i den tryckta ordboken eller var uppenbarliga synonymer som av någon anledning missats i den tryckta ordboken. I stort sett alla synonymer i den tryckta boken var med i Folkets synonymlexikon. De som saknades var mestadels omoderna eller mycket ovanliga.

6.2. Folkets lexikon

Det Lexinlexikon som flest använder och som antagligen därför genererar flest frågor till Språkrådet, som numera äger Lexin, är det svensk-engelska/engelsk-svenska. Det tillhör inte Språkrådets arbetsuppgifter att driva ett svensk-engelskt lexikon. Därför har Språkrådet nyligen överlåtit svensk-engelska/engelsk-svenska Lexin till KTH, som gör det tillgängligt i form av Folkets lexikon (FOLKETS). Projektet stöds ekonomiskt av .SE-stiftelsens Internetfond.

Tanken är att Folkets lexikon ska bli en fri resurs som ständigt utvecklas av folket. Forskare på Linköpings universitet har tagit fram en lång lista med översättningsförslag som bedöms av användarna av Folkets lexikon, se figur 4. På samma sätt som i Folkets synonymlexikon så kommer översättningsförslag, som användarna bedömer är bra, att läggas till lexikonet. Användare får också ge egna förslag som kommer att bedömas av andra användare och så småningom komma in i lexikonet. Även andra tillägg och ändringar kommer att bli möjliga att föreslå.



Figur 4: Användargränssnitt för Folkets lexikon

För att Folkets lexikon ska kunna växa ännu snabbare lägger vi in andra fria resurser i det, bland annat Folkets synonymlexikon och Svenskt associationslexikon (SALDO; se även Borins uppsats i denna volym).

7. Sammanfattning

Delar av femton års produktion från språkteknologigruppen på KTH har beskrivits. De flesta verktyg och resurser som tagits fram av gruppen är fritt tillgängliga och kan laddas ner från gruppens webbsida (HLTWEBB). En stor fördel med fritt tillgängliga verktyg är att såväl forskare vid högskolorna som utvecklare i industrin och privatpersoner som programmerar för nöjes skull kan få tillgång till moderna och fullskaliga verktyg och med hjälp av dem bygga nya språkteknologiska tillämpningar. Ett annat sätt att skapa stora fria resurser är med hjälp av crowdsourcing. Folkets synonymlexikon och Folkets lexikon blir hela tiden större och bättre, i och med att användare i tusental föreslår tillägg och bedömer varandras förslag. Det är ett alternativ till det traditionella mycket arbetskrävande sättet att ta fram lexikon. Min uppfattning är att crowdsourcing kommer att användas i många sammanhang i framtiden, men att det kommer att fungera väl och bli resultera i välfyllda lexikon endast i vissa gynnsamma fall. Popularitеten hos den webbsida där resursen byggs är kritisk, liksom att sidan måste besökas av användare som är villiga att hjälpa till att bygga just denna resurs.

Litteratur

Ordböcker

Klingspor, Richard: *Svenskt avstavningslexikon*. Tryckeriförlaget, Stockholm 1985.

SAOL 1986 = *Svenska Akademiens ordlista*, upplaga 11. Norstedts.

Annan litteratur

Bigert, Johnny 2005: Unsupervised evaluation of Swedish spell checker correction suggestions. I: *Nordiska datalingvistikdagarna 2005*, Joensuu, Finland.

Bloom, Burton H. 1970: Space/time trade-offs in hash coding with allowable errors. I: *Communications of the ACM*, volym 13, 422–426.

Damereau, Fred J. 1964: A technique for computer detection and correction of spelling errors. I: *Communications of the ACM*, volym 7, 649–664.

Hult, Ann-Kristin 2008: Användarna bakom loggfilerna. Redovisning av en webbenkät i Lexin online Svenska ord. I: *LexicoNordica*, volym 15, 73–91.

Kann, Viggo, Domeij, Rickard, Hollman, Joachim och Tillenius, Mikael 2001: Implementation aspects and applications of a spelling correction algorithm. I: L. Uhlirova, G. Wimmer, G. Altmann, R. Koehler: *Text as a Linguistic Paradigm: Levels, Constituents, Constructs*. Festschrift in honour of Ludek Hrebicek. Quantitative Linguistics, vol. 60, WVT, 108–123.

Kann, Viggo och Rosell, Magnus 2005: Free construction of a free Swedish dictionary of synonyms. I: S. Werner: *Nordiska datalingvistikdagarna 2005*, Joensuu, Finland, 105–110. <http://www.csc.kth.se/tcs/projects/infomat/rapporter/kann-rosell05.pdf>

Knutsson, Ola, Bigert, Johnny och Kann, Viggo 2003: A robust shallow parser for Swedish. I: *Nordiska datalingvistikdagarna 2003*, Reykjavik, Island. <http://www.nada.kth.se/theory/projects/xcheck/rapporter/gta03.pdf>

Sjöbergh, Jonas och Kann, Viggo 2006: Vad kan statistik avslöja om svenska sammansättningar? I: *Språk och stil*, volym 16, 199–214.

Internethänvisningar

AVSTAV = Språkrådet: *Avstavningslexikon*. http://avstava.appspot.com

FOLKETS = Hollman, Joachim och Kann, Viggo: *Folkets lexikon*. <http://folkets-lexikon.csc.kth.se>

GRIM = Westlund, Knutsson och Sjöbergh med flera: *Grim – en interaktiv miljö med fokus på det svenska språket*. <http://skrutten.nada.kth.se/grim/>

HLTWEBB = Språkteknologi vid KTH. <http://www.csc.kth.se/theory/humanlang/>

LEXIN = Språkrådet: *Lexin*. <http://lexin.nada.kth.se>

SALDO = Språkbanken: Saldo*, svenskt associationslexikon*. <http://spraakbanken.gu.se/saldo/>

SKANDORD = Nordiska språksekretariatet: *Skandinavisk ordbok*. <http://www.nada.kth.se/skandlexikon/>

Stallman, Richard 1996: *The Free Software Definition*. <http://www.gnu.org/philosophy/free-sw.html>

STAVA = Kann, Viggo: *Stavningskontrollprogrammet Stava*. KTH. <http://www.nada.kth.se/stava>

SYNLEX = Kann, Viggo och Rosell, Magnus: *Folkets synonymlexikon*. <http://lexin.nada.kth.se/synlex.html>

TAGGSTAVA = Kann, Viggo: *Ordklasstaggaren Taggstava*. <http://www.csc.kth.se/tcs/projects/granska/taggstava.html>

TVÄRSLÅ = Projektet Nordisk nätordbok: *Tvärslå*.
<http://ordbok.nada.kth.se>

Viggo Kann
professor
KTH Skolan för datavetenskap
och kommunikation
SE-100 44 Stockholm
viggo@kth.se

FinnWordNet – WordNet på finska via översättning

Krister Lindén & Lauri Carlson

FinnWordNet is a WordNet for Finnish that conforms to the framework given in Fellbaum (1998) and Vossen (1998). FinnWordNet[1] is open source and currently contains 117,000 synsets. A classic WordNet consists of synsets, or sets of partial synonyms whose shared meaning is described and exemplified by a gloss, a common part of speech and a hyperonym. Synsets in a WordNet are arranged in hierarchical partial orderings according to semantic relations like hyponymy/hyperonymy. Together the gloss, part of speech and hyperonym fix the meaning of a word and constrain the possible translations of a word in a given synset. The Finnish group has opted for translating Princeton WordNet 3.0 synsets wholesale into Finnish by professional translators, because the translation process can be controlled with regard to quality, coverage, cost and speed of translation. The project was financed by FIN-CLARIN at the University of Helsinki. According to our preliminary evaluation, the translation process was diligent and the quality is on a par with the original Princeton WordNet.

1. Inledning

Ett WordNet är en tesaurus som består av grupper av synonymer, dvs. ord som hör till samma ordklass och som är utbytbara i en given kontext. Sådana grupper av synonymer kallas även *synset*. I WordNet är synonymgrupperna partiellt hierarkiskt ordnade enligt semantiska relationer såsom hyperonymer, hyponymer, meronymer, antonymer, osv. Varje synset (med några få undantag) har en förklaring som exemplifierar eller beskriver dess betydelse.

1 FinnWordNet: http://www.ling.helsinki.fi/cgi-bin/fiwn/search

En typisk WordNet-förklaring inkluderar en ordboksdefinition och ett exempel, t.ex. *devilize* 'förvandla till en djävul eller göra demonisk; "en människa demoniserad av kriget"'. Somliga förklaringar innehåller bara en hyperonym (exempelvis ett namn på en djurart).

FinnWordNet är ett WordNet för finska som följer den struktur som beskrivs i Fellbaum (1998) och Vossen (1998) för Princeton WordNet, som är öppen källkod och för närvarande innehåller 117 000 synset. FinnWordNet är en översättning av synseten i Princeton WordNet 3.0.

Härnäst ger vi en överblick över existerande WordNet på andra språk. Längre fram går vi igenom fördelarna med manuell översättning, arbetsflödet och några teoretiska och praktiska problem som vi stötte på innan vi utvärderar arbetet och nämner något av det som återstår nu när grovjobbet är gjort.

2. WordNet

WordNet på andra språk har skapats med olika metoder. Vi börjar med en kort genomgång av huvudalternativen. Dessutom försöker vi uppskatta hur mycket arbete det innebär att göra ett nytt WordNet utgående från Princeton WordNet.

2.1. WordNet på andra språk

WordNet har också skapats för andra språk än engelska, men generellt sett har de inte lika god täckning som Princeton WordNet. I tabell 1 ser vi en jämförelse av storlekarna på WordNet för olika språk och språkfamiljer. Siffrorna har hämtats från källor på internet under våren 2010.

WordNet	Synset
Princeton WordNet (Fellbaum 1998, WordNet 3.0)	~120 000
EuroWordNet (Vossen 2004a)	~10 000–50 000
BalkaNet (Vossen 2004b)	~18 000
Polskt WordNet (Piasecki m.fl. 2009)	~18 000
Danskt WordNet (DanNet 2009, Pedersen 2010)	~41 000–60 000
Svenskt WordNet (Viberg m.fl. 2002)	~15 000
NorNett (Fjeld och Nygaard 2009)	~50 000 synset? (80 000 relationer)

Tabell 1: WordNet på olika språk och för olika språkfamiljer och deras storlekar i antal synset

2.2. Att skapa ett WordNet för ett nytt språk

Om man vill skapa ett WordNet för ett nytt språk, kan man välja mellan tre olika alternativ: skapa ett WordNet från grunden, översätta ett annat WordNet eller använda en toppontologi och utvidga den med en lokal synonymordbok.

Om man väljer det första alternativet och skapar ett WordNet från grunden, måste synseten utvinnas med olika automatiska metoder ur stora korpusar. Ett bra exempel på detta är polska (Piasecki m.fl. 2009), där synseten utvanns från polska korpusar. För att säkerställa att de mest väsentliga orden och deras synonymer kom med, blev den polska gruppen ändå tvungen att använda en ordlista över det centrala ordförrådet i polska med cirka 30 000 ord. Dessutom måste synseten i ett WordNet som skapas från grunden separat länkas till synseten i ett annat WordNet om man vill skapa en tvåspråkig ordbok.

Det andra alternativet, dvs. att översätta t.ex. Princeton WordNet rubb och stubb, tonar ned argumentet att varje språk är så

annorlunda att man måste skapa alldeles egna synonymgrupper och synonymhierarkier för varje språk. De flesta ord i ett språk är namn på objekt och fenomen i en för mänskligheten gemensam extern verklighet som varje språk har hittat på eller lånat namn för. Det kan finnas en viss skillnad i hur små nyanser som fått egna ord beroende på hur viktigt ett fenomen är inom ett språkområde, men i stort är den grundläggande fysiska verkligheten densamma i alla kulturkretsar: solen går varje dag upp i öster, vattnet är vått och rinner, vi föds, blir hungriga, äter mat och dör, osv. För abstrakta begrepp kan man däremot förvänta sig en större divergens. Ett exempel i Princeton WordNet är specifikt amerikanska begrepp med alla sina moderna sidobetydelser och nyanser som inte finns i ett potentiellt målspråk. Se exempel i tabell 2.

engelska	**svenska**	**finska**
hungry	hungrig	nälkäinen
hunger	hungra	nähdä nälkää
Thanksgiving	tacksägelsedag?, skördefest?	kiitospäivä, ≠ kekri

Tabell 2: Exempel på likheter och skillnader mellan konkreta och kulturbundna begrepp i olika språk. Hungern är global men festerna varierar.

Ett tredje alternativ och en medelväg att skapa WordNet för olika språk är därför att översätta 5000 centrala begrepp i Princeton WordNet för att skapa en kärna som sedan utvidgas med en lokal synonymordbok. Detta är den väg man har tänkt sig för DanNet, där man börjat med att skapa den lokala synonymordboken på basen av en lokal ordbok (Pedersen 2010). Avsikten är att sedan länka DanNet till Princeton WordNet. Detta skapar en länk mellan de mest centrala begreppen i de båda språken och underlättar användningen av den nya ordboken vid översättning. Detta användes i t.ex. EuroWordNet (Vossen 2004a). Världsorganisatio-

nen för WordNet, The Global WordNet Association (GWA), har gjort ett urval av 5000 grundläggande begrepp enligt strukturella principer för WordNet i olika språk såsom begreppens position i betydelsehierarkin, antal relationer, frekvens med vilken de förekommer i definitioner eller förklaringar och vilka deras morfologiska komplexitet och beroenden är. I många fall finns det ingen statistik på hur vanlig en viss betydelse är och därför är den lista på 5000 grundbegrepp som GWA föreslår en rätt brokig samling med det amerikanska TV-programmet *60 minutes* men utan ordet *television.*

Å ena sidan fordrar skapandet av en enspråkig resurs från grunden att man upprepar årtionden av lexikografiskt arbete som investerats i Princeton WordNet, men å andra sidan kräver en översättning av hela Princeton WordNet att man även tar ställning till några specifikt amerikanska fenomen som inte nödvändigtvis har en exakt motsvarighet i målspråket. I många språk finns dessutom existerande synonymordböcker och därför har många WordNet-projekt valt den tredje vägen, dvs. en kombination av översättning och en lokalt skapad betydelsehierarki.

Vi anser att den grundläggande enheten i översättning är en betydelseenhet som kan representeras av ett synset med hyperonym, ordklass och förklaring som tilläggsinformation för att styra översättningen av de olika synonymerna i ett synset. Vi valde därför att översätta alla synset. Detta ger oss en parallell uppsättning av synset i de två språken så att vi kan återanvända de flesta semantiska relationer som definierats i Princeton WordNet och dessutom få en tvåspråkig ordbok.

Finska är som språk obesläktat med de andra språken med offentligt tillgängliga WordNet. Dessutom är vi inte medvetna om att alla synset i Princeton WordNet skulle ha översatts tidigare. Det återstår att utvärdera i ett senare skede hur många specifikt finska ord som saknas och hur många specifikt finska fenomen och betydelser som behöver läggas till.

2.3. Princeton WordNet 3.0

Nedan i tabell 3 och 4 ger vi en uppskattning på mängden av det översättningsarbete som vi valt. I tabell 3 ser vi att Princeton WordNet 3.0 har 117 659 synset eller betydelseenheter. Det finns 155 287 olika ord (strängar) i dessa synset. Några ord förekommer i flera synset. Detta skapar totalt 206 941 synonymer eller ord-synset-par att översätta.

Princeton WordNet	ord	synset	synonymer
substantiv	117 798	82 115	146 312
verb	11 529	13 767	25 047
adjektiv	21 479	18 156	30 002
adverb	4 481	3 621	5 580
Totalt	**155 287**	**117 659**	**206 941**

Tabell 3: Antal ord, synset och synonymer i Princeton WordNet

I tabell 4 ser vi att de flesta ord bara har en betydelse, dvs. de förekommer i bara ett synset. Eftersom det finns färre synset än ord är det naturligt att det finns ett antal synset som har fler än ett ord i sig. I medeltal är antalet synonymer per synset 1,8. Dessutom förekommer somliga ord (strängar) i flera olika synset, dvs. är homonymer för olika betydelser. I medeltal har ett ord 1,3 betydelser. De flesta ord har bara en betydelse, så om man betraktar bara de ord som har flera betydelser, har de flertydiga orden nästan 3 betydelser i medeltal.

Princeton WordNet	monosema ord	homonyma ord	homonymer totalt
substantiv	101 863	15 935	44 449
verb	6 277	5 252	18 770
adjektiv	16 503	4 976	14 399
adverb	3 748	733	1 832
Totalt	**128 391**	**26 896**	**79 450**

Tabell 4: Antal monosema och homonyma ord samt det totala antalet homonymer i Princeton WordNet

3. Manuell översättning av synset

Vi valde professionell manuell översättning av alla synset i Princeton WordNet till finska eftersom den manuella översättningsprocessen är kontrollerad i flera avseenden:

1. Manuell översättning görs av människor och garanterar därför *hög kvalitet* i jämförelse med maskinöversättning och de datautvinningstekniker som normalt används för att upptäcka ords betydelser i stora korpusar.
2. Översättning av alla synset garanterar att vi får en *stor mängd*, dvs. cirka 200 000 betydelser, så att vi får mer än ytterligare en kärnvokabulär med bara de mest centrala ordbetydelserna.
3. Att använda en översättningsbyrå till ett överenskommet pris per ord garanterar en *kontrollerad kostnad*.
4. En översättningsbyrå garanterar även en viss *hastighet*, dvs. 3–6 månader för en första version med upp till 4–5 översättare som jobbar parallellt.

I det följande går vi in på ytterligare några fördelar med ett översatt WordNet och betraktar specifika utmaningar när det gäller att

översätta WordNet jämfört med att översätta vanliga dokument. Vi ger en överblick över översättningsprocessens arbetsflöde och kvalitetskontroll samt bekantar oss med de metadata som översättaren förväntas bidra med för att hjälpa till med efterbearbetningen för att förbättra kvaliteten.

3.1. Ytterligare fördelar med ett tvåspråkigt WordNet

Översättning gav oss möjlighet att kontrollera processen och producera resultat effektivt, men dessutom ville vi skapa en tvåspråkig resurs som är direkt länkad till Princeton WordNet. Vi skapade ett fritt tillgängligt tvåspråkigt finsk-engelskt WordNet på fyra månader och hoppas att det kan fungera som ett exempel för andra språk.

Den tvåspråkiga ordboken bestående av två parallella WordNet kan användas till informationssökning och med lite modifikation även för maskinöversättning tack vare det stora lexikonet med 100 000–200 000 översättningar med disambiguerad betydelse. Det bör påpekas att nyckelfraser i informationssökning inte är de vanligaste orden utan de mest specifika orden som namn och termer, som ofta saknas i enkla tvåspråkiga ordböcker med enbart kärnvokabulär. Det bör också påpekas att översättningar utan kontext inte är direkt användbara i maskinöversättning, men i WordNet fungerar de engelska förklaringarna och definitionerna som en grund för att utvinna en kontext.

Vi tror att det finns tillräckliga likheter mellan engelska och finska lexikala enheter för att överbrygga eventuella olikheter, även om en grundlig utvärdering är nödvändig för att ge en slutlig indikation på hur många specifikt finska kärnbetydelser och kärnord som saknas och hur mycket manuell och halvautomatisk bearbetning som ännu behövs.

3.2. Översättningsutmaningar för WordNet

Vårt dilemma är att vi vill ha alla tänkbara synonymer för en specifik ordbetydelse men samtidigt vill vi ha enbart de relevanta synonymerna. Vårt mål står delvis i konflikt med hur professionella översättare tränas att översätta kommersiella dokument, dvs. att i mån av möjlighet undvika icke-standardiserade ordvarianter. Nu vill vi ha alla varianter i tillägg till standardöversättningen. Vi löste problemet genom att be översättarna hitta finska översättningar till ett engelskt synset som helhet men att översätta varje engelsk synonym i ett synset med sina närmaste särskiljande översättningar. Detta uppmuntrade översättare att leta efter mer än en finsk motsvarighet för att matcha de engelska synonymerna. För att underlätta och stabilisera översättningen bad vi översättaren fästa uppmärksamhet vid[2]:

1. *förklaring*, dvs. betydelsekontext eller betydelsedefinition
2. ordets *hyperonym*, dvs. begreppets abstraktionsnivå
3. *ordklass*, dvs. den morfo-syntaktiska kontexten

Vi använde en vanlig översättningseditor som stöder XML-format för att skydda de fält som vi ville visa översättarna enbart som tilläggsinformation för att guida deras översättning av synseten.

3.3. Arbetsflöde och kvalitetskontroll

Först tänkte vi dela in WordNet i sektioner enligt ämnesområde, men alla ord har inte en angivelse om ämnesområde och många ämnesområden hade relativt få ord, så vi övergav idén till förmån för en mera genomgripande och lättgenomförbar princip. Vi an-

2 De här tre parametrarna hade i förberedande experiment med studerande i översättning konstaterats vara värdefulla stabiliserande riktlinjer för översättning av synset i WordNet.

vände följande arbetsflöde för att åstadkomma kvalitetskontroll och en någotsånär jämn fördelning av arbetet mellan översättarna:

1. WordNet transformerades till XML-format och delades in i 20 alfabetiskt ordnade synsetsamlingar med enbart de fält som var väsentliga för översättningsarbetet. Cirka 10 000 ortografiskt identifierade egennamns[3] synset (med 20 000 varianter) åtskiljdes i en separat egennamnsfil. Många mindre självklara namngivna objekt och fenomen blev trots det kvar i de övriga filerna.
2. En synsetsamling överlämnades till en översättarkandidat som först producerade ett översättningssampel för 1000 engelska synonymer.
3. Översättningsprovet inspekterades av en översättare ansvarig för kvalitetskontrollen.
4. Om det inte fanns några större klagomål, gavs översättaren ytterligare instruktioner på basis av översättningsprovet och han fick hela filen att översätta inom en viss tid.
5. Efter att filen överlämnats, gjordes en slutlig kvalitetskontroll för att bestämma om översättningen föll inom ramen för de uppställda kvalitetskraven.

3.4. Praktiskt översättningsarbete

För att läsaren bättre skall förstå översättningsprocessen visar vi ett förenklat sampel av en XML-kodad fil:

```
<GLOSS>an Egyptian descended from the ancient
Egyptians
<HYPER ID="109700492">Egyptian
<SYNONYM PoS="n">
```

3 Egennamn, dvs. namn på personer och fenomen, har separata synset i WordNet. Synsetets förklaring kan vara väsentlig för att identifiera vem eller vad det är fråga om. Namn på växter och djur har normalt flera varianter bl.a. ett latinskt namn.

```
<Tuv Lang="EN-US">Copt
<Tuv Lang="FI">koptilainen
```

Vi valde att använda SDL Trados som vår översättningsomgivning, eftersom den är vanlig bland professionella översättare. Trados TagEditor stöder översättning av XML-dokument. Trados håller översatta ord i översättningsminnet. Ett översättningsminne erbjuder mycket lite hjälp när man skall översätta synset där en ny förekomst av ett ord i ett annat synset sannolikt kräver en annan översättning. Editorn hjälper till att begränsa editeringen till de fält som skall översättas och tillåter oss att definiera de specifika koder som skall användas under översättningsprojektet. Översättarna använder koderna för att ange metadata om sina översättningar för att styra kvalitetskontrollen och vidareutvecklingen till ställen där det kan finnas brister. Efter en diskussion med översättarna beslöt vi oss för följande koder:

- `<or/>`
- `<approximate/>`
- `<broader/>`
- `<narrower/>`
- `<unconfirmed/>`
- `<note> ... </note>`
- `<GEN/>`

Med `<or/>` separerar vi alternativa översättningar av samma term för att undvika tvetydighet mellan objektspråk och metaspråk (interpunktion såsom komma kan förekomma som en del av en objektspråksfras). Med `<approximate/>` kan översättaren ange att han efter en rimlig insats inte kunde hitta någon bra motsvarighet. Med `<broader/>` indikerar han att översättningen är mera generell än originaltermen och med `<narrower/>` att översättningen är mera specifik än originaltermen. Avsikten med koderna var inte att problematisera varje motsvarighet utan att låta översättaren

ange lexikala luckor. I praktiken användes koderna rätt lite (cirka en gång per tusen betydelser). Exempelvis användes `<broader/>` för översättning med hyperonymen *nahka* 'leather' för *ooze leather,* och `<narrower/>` för översättning med hyponymen *yläpurentainen* 'overshot (jaw)' för *overshot.* Koden `<unconfirmed/>` användes om översättaren ansåg att termen var korrekt men inte kunde bekräftas i skrivna källor. Koden `<note>` ... `</note>` tillät översättaren att lägga till den information han ansåg lämplig för senare inspektion och vidareutveckling av översättningen, t.ex. för att indikera problem i den engelska källan.

Översättarna fick en manual före uppgiften med exempel på goda översättningar och sådana översättningar som borde undvikas. Kvalitetskontrollanten kunde därför närmast hänvisa till manualen för att ge ytterligare instruktioner. Översättarna ombads tänka ut motsvarigheter som kunde ersätta källtermen i synsetkontexten, undvika nonsensöversättningar, parafraser och betydelseförklaringar. Vidare ombads de undvika sammanblandning av objektspråk och metaspråk såsom parentetiska tillägg, ad hoc grammatikaliska förklaringar, osv.

Översättarna instruerades att bevara samma ordklass som i källspråket när det var möjligt. Avvikelser från detta borde anges för att underlätta en senare automatisk morfologisk analys av översättningen.

Koden `<GEN/>` användes i det allmänt förekommande fallet att ett engelskt adjektiv översattes med genitiv av ett finskt substantiv. En mera komplett annotering av grammatiken kunde ha utvecklats men det avstod vi avsiktligt från för att inte komplicera översättarens arbete. En ytterligare genomgång av materialet av en tränad lingvist kan bättre åtgärda den aspekten.

4. Översättningsproblem

Teoretiskt sett finns det många problem förknippade med översättningen av en tesaurus. Vi tittar först på ett prototypexempel med icke-matchande semantiska fält, för att sedan titta på vilka problem som förekom i praktiken. Vi går också igenom hur vi kontrollerade översättningskvaliteten under översättningsprocessen och gör några allmänna observationer om den översättningskvalitet vi fick.

4.1. Teoretisk utmaning – icke-matchande semantiska fält

Från en teoretisk synvinkel har det största översättningsproblemet beskrivits som icke-matchande semantiska fält mellan begreppen i två språk. Beträffande WordNet i synnerhet kan vi ha icke-matchande synset i två språk, eftersom inte alla ord som är synonymer i ett språk har ekvivalenter som uppfattas som synonymer i ett annat språk, om man betraktar ordens kärnbetydelse eller mest frekventa betydelse. Vår intuition är att detta vanligen inte är ett problem vid översättning av WordNet, eftersom ordklass, förklaring och hyperonym ger en klar indikation om vilken översättning som lämpar sig i den givna kontexten. Dessutom har den semantiska rymden i Princeton WordNet under en lång tid granskats ur många synvinklar.

Låt oss ta en titt på ett exempel som har icke-matchande semantiska fält i engelska och finska på grund av det sätt som engelska och finska hanterar nationalitetsord. I båda språken är nationaliteten ofta ett adjektiv härlett från namnet på landet. På finska sammanfaller namnet på språket ofta med landet. På engelska däremot sammanfaller namnet på språket vanligen med nationaliteten.

```
<GLOSS> an Egyptian descended from the ancient Egyptians
<HYPER> Egyptian
<PoS="n" SENSE="1" COUNT="0">
   "EN-US": Copt
   "FI": koptilainen <or/> kopti

<GLOSS> the liturgical language of the Coptic Church
used in Egypt and Ethiopia; written in the Greek al-
phabet
<HYPER> Egyptian
<PoS="n" SENSE="1" COUNT="0">
   "EN-US": Coptic
   "FI": koptin kieli <or/> kopti

<GLOSS> of or relating to the Copts or their Church or
language or art; "the distinctive Coptic art of 6th-
century Christian Egypt"
<HYPER> Egyptian
<PoS="a" SENSE="1" COUNT="0">
   "EN-US": Coptic
   "FI": koptilainen
```

I detta fall är WordNet tillräckligt finmaskigt för att översättningen av begreppen såsom de har underindelats i WordNet inte skall ställa till problem. Detta kan förstås bero på att det är fråga om en hel grupp av ord som har liknande icke-matchande fält i många språk så problemen har för länge sedan upptäckts och åtgärdats. I exemplet kompliceras översättningen av att ordet även råkar vara en benämning på en religiös inriktning, men under hyperonymen *Church* finns det även ett separat synset för *Coptic Church* som då översätts med *koptilaisuus*.

Vi ser det som en styrka hos Princeton WordNet att det har utvecklats och uppdaterats lexikografiskt under många år. Detta arbete tas tillvara vid översättning. Dock återstår vissa problemområden vid närmare granskning.

Ett framträdande problem är behandlingen av idiom i Word-

Net. Ibland förekommer idiomen exemplariskt dokumenterade som separata betydelseenheter med egna synonymer – dock inte alltid, t.ex. finns *kick the bucket* som underbetydelse av *kick*. I störande många fall låter WordNet en icke-kompositionell betydelseenhet utgöra ett separat synset, t.ex. *tartaric* i *tartaric acid* eller *take* i *take in*. Detta är ett återfall till det traditionellt lemma-baserade tänkandet i pappersordböcker där lexikografen även förutsattes hitta ett lämpligt uppslagsord för flerordiga uttryck. Sökningen sköter datorn numera alldeles utmärkt, så lexikografen kan koncentrera sig på innehåll och betydelse. Det bästa en översättare kan göra är att ändå översätta hela idiomet och anteckna svårigheten i en kommentar.

4.2. Problem i praktiken

I praktiken visade det sig att det ibland är svårt att hitta rätt eller ens någon översättning alls för:

- medicinska termer såsom varunamn på mediciner
- växter och organismer specifika för Amerika
- kemiska substanser såsom varunamn på aktiva ingredienser
- olika ismer såsom religiösa rörelser
- lagtekniska termer
- affärstermer
- andra termer som är specifika för amerikansk kultur

Ett exempel på icke-översättbar amerikanska är *hanging chad*, dvs. en icke-fullföljd perforering av en röstsedel på grund av en felaktig röstningsmaskin, vilket var förstasidesnyheter under valåret 2000 i USA. Det finns även specifika WordNet-ord som är svåra att hitta på webben utanför själva WordNet (och dess kloner). Ett exempel är *spiritize* 'att förse med ande eller andlighet'. Andra exempel är *spouter* 'en frustande val', *tapper* 'en person som slår lätt, ofta upp-

repat, på en yta', och *Tareekh e Kasas* 'en organisation av muslimer i Indien som dödade hinduer i september 2002 och som tros ha samröre med muslimska terrorister i Pakistan'. En uppdaterad version av WordNet på nätet kunde kanske göra upp användningsstatistik på orden om man vill ta bort mera efemära dagsländor och *hapax legomena*. Å andra sidan är de här orden mest ett problem vid översättning av WordNet, vilket ju görs rätt sällan. I daglig användning av Wordnet söker ingen upp de udda orden och därför stör de inte heller. I praktiken kan det kosta mer att ta bort dem än att betala för lite mer dataminne.

De udda orden är exempel på kulturbundna fenomen eller *realia* som har utvecklats i det amerikanska samhället under de senaste tvåhundra åren och som därför är specifika för amerikansk kultur. Många är namn på ideologier och sociala fenomen som är okända i finländsk kultur och i det finländska samhället. Detta verkar bekräfta vår ursprungliga hypotes att bara relativt nya och kulturspecifika fenomen orsakar verkliga problem vid översättning. Lyckligtvis har de problematiska orden sällan synonymer. Dessutom är de en liten minoritet av alla ord i WordNet och motiverar inte varför man skulle göra om 95 % av de icke-problematiska synseten från grunden.

En liten motsatt brist i WordNet är att det ibland görs distinktioner där samma ord används i flera olika synset utan att det har betydelse vid översättningen, dvs. alla nyanserna har samma ord även i målspråket. Ibland kan detta bero på en tillfällig likhet mellan språken.

4.3. Indikatorer på översättningskvalitet

För att kontrollera översättningskvaliteten skapade vi olika indikatorer. Vi följde upp varje indikator genom att jämföra en översättares produktion med medeltalet för alla översättare. En stor avvikelse fäster kvalitetskontrollantens uppmärksamhet på ett spe-

cifikt synset. Förutom statistiken får kvalitetskontrollanten exempel på avvikande synset sorterade i sjunkande grad av avvikelse.

De två huvudsakliga indikatorerna är antalet olika översättningar per synset, och antalet oöversatta eller identiskt översatta ord. Andra indikatorer följer översättarnas användning av annotationskoder.

Indikatorn för antalet olika översättningar per synset anger hur många synonymer det finns i ett synset på finska jämfört med antalet synonymer på engelska. Man kan använda denna indikator för att upptäcka den variationsbrist som uppstår t.ex. om en översättare använder samma eller få översättningar för alla synonymer i ett synset.

Vi tillåter att ett ord efter ett lämpligt övervägande kanske inte har någon känd finsk översättning, vilket kan signaleras genom att kopiera den engelska strängen som sådan. Vi följde användningen av denna icke-översättning med en indikator på hur många strängar som är identiska på engelska och finska. I teorin borde det vara tillämpbart endast på sådana namn som inte behöver translittereras men för ett litet förbestämt pris per ord kan en översättare inte förväntas använda många dagar på gediget terminologiarbete. I stället skall de indikera sin osäkerhet med en lämplig kod, så att vi kan återkomma till detta ställe i ett senare skede. Om procenten oöversatta ord är hög för en översättare måste kvalitetskontrollanten ge filen extra uppmärksamhet för att uppdaga orsaken.

Liknande indikatorer för hur ofta koder används i medeltal gjordes också upp. Ett överdådigt antal varianter syns i en onödigt stor användning av `<or/>`. För många varianter är ibland en indikation på att översättaren är osäker på den korrekta varianten och garderar sig genom att generera flera möjligheter. Det samma gäller även för onödigt liberala översättningar vilket syns i att koderna `<approximate/>` och `<unconfirmed/>` används mer än väntat.

Inom översättningsverksamhet är det praxis att kvalitetskont-

roll får ta högst 20 % av översättningstiden. För att höja hastigheten på granskningen, beräknade vi processindikatorerna som medeltal på alla synsetsamlingar för att bestämma väntevärden för olika koder samt för storleken på synset för engelska och finska. Synsetsamlingar där indikatorerna avvek från medeltalet granskades mera utförligt. Alla synset jämfördes med medeltalet för sina synsetsamlingar. De 500 mest avvikande synseten i en samling inspekterades ytterligare, men vi kunde inte upptäcka några försök att medvetet missbruka processen. De misstag som kunde hänföras till enstaka tillfällen av bristande uppmärksamhet rapporterades som generell feedback till alla översättare.

4.4. Observationer angående den levererade översättningskvaliteten

Några generella observationer om översättningsprocessen kan göras. Först och främst bör vi notera att översättarna var samvetsgranna och gjorde sitt bästa som professionella översättare.

I början tenderade några översättare att vara alltför försiktiga och oftare använda koderna `<unconfirmed/>` och `<approximate/>`, men användningen minskade mot slutet. En annan strategi för att dölja tillfällig osäkerhet i början var att hitta på oproportionerligt många synonymer. För andra än egennamn, kopierades mindre än 10 % av strängarna direkt. Oftast kopierades den engelska strängen till finska för egennamn, varunamn eller andra namngivna enheter som inte behövde translittereras till finska.

I början försökte översättarna generera tämligen kompletta grupper av synonymer. Mot slutet av sina synsetsamlingar tenderade de ändå att falla tillbaka på den normala rutinen att förse varje synonym i synsetet med bara en översättning. Vi kan bara spekulera i orsaken till detta men en är säkert behovet att ta igen förlorad tid om alltför mycket tid användes på början av filen.

Detta kom inte som någon överraskning eftersom vi egentligen

inte räknade med att få flera översättningar per synonym i synsetet. Översättarna fick betalt för att översätta varje synonym minst en gång och vi antog att detta borde ge oss en tillräcklig mängd betydelser per ord och synset.

Även om vi först nyligen har påbörjat en grundligare evaluering av resultatet, kan vi erbjuda några preliminära siffror. I tabellen inbegriper *finska* endast icke-identiska översättningar till finska (vilket utesluter ca 26 000 egennamn och andra identiskt översatta strängar som finns med i siffrorna för engelska).

Jämförelsetal	**engelska**	**finska**
synonymer (**token**)	206 723	181 753
ord (**typer**)	148 556	99 639
synonymer / ord	1,39	1,81
synonymer: eng./fin.	1,15	
ord: eng./fin.	1,50	

Tabell 5: Jämförelsetal för engelska och finska synonymer per ord

Vi tolkar siffrorna i tabell 5 på följande sätt. De 181 753 finska synonymerna utgör översättningar till 206 723 engelska synonymer. Skillnaden i antal består mest av ovan nämnda identiskt översatta engelska namn och andra oöversättbara ord. Endast 287 synonymer förblev helt oöversatta. De flesta av dem var siffror som översättningsredskapet inte tillät översättas.

De 99 639 olika finska orden översätter 148 556 olika engelska ord. De engelska siffrorna inkluderar identiskt översatta strängar. När deras antal dras av från antalet engelska ord i jämförelsen tyder siffrorna på att minskningen i lexikal variation är liten, dvs. översättarna har genererat ungefär lika många olika och specialiserade ord på finska som det fanns olika ord i den engelska förlagan. Det är mest fördelningen som är olika.

Om man jämför fördelningen av ord på synset i det engelska

WordNet med fördelningen i översättningen, var de finska översättningarna 1,3 gånger mer flertydiga än det engelska originalet med 1,81 betydelser per finskt ord i medeltal jämfört med originalets 1,39 betydelser per engelskt ord. Om man även här lägger till de ca 26 000 identiskt översatta strängarna till antalet finska ord är antalet betydelser per ord bara 1,61 på finska mot 1,39 på engelska. Man kan dra slutsatsen att något fler finska än engelska ord används i två eller flera synset.

För att ge ett exempel på den lexikala variationen på finska, jämför vi synonymerna för *wave* på engelska med synonymerna för *aalto* på finska. I WordNet är *wave* en hyponym till *motion.* Ordet *wave* får två direkta finska översättningar: *aalto* och *laine*. Bland sina WordNet-hyponymer har *wave* ord såsom *billow, surge, comber, fluctuation, ripple, rippling, riffle, wavelet, roll, roller, rolling wave, seiche, surf, breaker, swell, crestless wave, whitecap, white horse.* Dessa 18 engelska ord fick 10 finska översättningar (inte uppräknade här). En jämförelse med en finsk synonymordbok på nätet ger ytterligare nio synonymer. Av dessa nio förekommer sex i FinnWordNet som översättningar av ord under närliggande men andra hyperonymer: *spray, breeze, wind, flow, foam*. Man kan argumentera för att några av dem hör till begreppet 'vågor' men metonymt såsom *pärske* 'spume'. Bland de föreslagna orden i synonymordboken hade de poetiska orden såsom *pärsky*, *lakkapää* and *vaahtopää* 'whitecap' inte alls nämnts av översättarna (*whitecap* hade översatts med *vaahtopäinen aalto* 'a white-capped wave'). Å andra sidan har WordNet översättningarna en del faktaorienterade hyponymer till *wave* som inte finns med i synonymordboken såsom *tsunami* och *backwash*.

5. Sammanfattning och fortsättning

Tidtabellen för att skapa en första version av FinnWordNet var strikt. Översättningsprocessen började i november 2009 och råöversättningen var klar i slutet av mars 2010. Det tog alltså 100 arbetsdagar. Om översättarna hade översatt med en jämn hastighet på tusen synonymer om dagen, skulle detta grovt räknat ha inneburit två heltidsöversättare. I praktiken var genomsnittshastigheten mellan 500 och 1000 betydelser per dag för en översättare.

Behovet att förbättra FinnWordNet kommer att utvärderas genom att titta på vilka finska grundformer som inte förekommer i FinnWordNet med speciell vikt på saknade högfrekventa ord. Dessutom behöver vi identifiera saknade betydelser av existerande grundformer, vilket antagligen måste upptäckas med halvautomatiska metoder i stora korpusar så som i arbetet med det polska WordNet (Piasecki m.fl. 2009). En separat studie kommer att göras manuellt för att bestämma om det finns delar av det översatta WordNet som inte har en adekvat ontologisk struktur för finska.

För att korrigera misstag i översättningarna och lägga till nya översättningar, kommer den finsk-engelska WordNet-ordboken att göras offentligt tillgänglig på nätet med en möjlighet även för allmänheten att föreslå korrigeringar och förbättringar.

Litteratur

DanNet 2009 = *DanNet – det danske wordnet*. http://wordnet.dk/ (mars 2010)

Fellbaum, Christiane (red.) 1998: *WordNet: An Electronic Lexical Database.* Cambridge/London/England: The MIT Press.

Fjeld, Ruth V. och Lars Nygaard 2009: NorNett — a monolingual wordnet of modern Norwegian. I: *NEALT Proceedings Series* 7, 13-16. http://hdl.handle.net/10062/9837

Pedersen, Bolette Sandford 2010: Semantiske sprogressourcer – mellem sprogteknologi og leksikografi. I: *LexicoNordica 17* (i dette bind).

Piasecki, Maciej m.fl. 2009: *A Wordnet from the Ground Up.* Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Viberg, Åke m.fl. 2002: *The Swedish WordNet Project*. http://www.lingfil.uu.se/personal/viberg/SwedishWordNet2.pdf (mars 2010)

Vossen, Piek (red.) 1998: *EuroWordNet: a multilingual database with lexical semantic networks for European Languages.* Dordrecht: Kluwer.

Vossen, Piek 2004a: EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an Inter-Lingual-Index. I: *International Journal of Linguistics 17(2)*, 161-173

Vossen, Piek 2004b: Introduction. I: *Romanian Journal of Information Science and Technology 7(1-2)*, 5-6.

WordNet 3.0 = *WordNet – A lexical database for English*. http://wordnet.princeton.edu/ (mars 2010)

Krister Lindén
docent, FD
Helsingfors universitet
Unionsgatan 40
FIN-00014 Helsingfors universitet
krister.linden@helsinki.fi

Lauri Carlson
professor, FD
Helsingfors universitet
Unionsgatan 40
FIN-00014 Helsingfors universitet
lauri.carlson@helsinki.fi

Lexicon Acquisition through Noun Clustering

Anna Björk Nikulásdóttir & Matthew Whelpton

This paper describes an experiment with clustering of Icelandic nouns based on semantic relatedness. This work is part of a larger project aiming at semi-automatically constructing a semantic database for Icelandic language technology. The harvested semantic clusters also provide valuable information for traditional lexicography.

1. Introduction

Semantic resources are already an established part of natural language processing (NLP) applications for dominant languages. Following the Princeton WordNet (Fellbaum 1998) for English, many other languages have created their own WordNet-like resources (cf. http://www.globalwordnet.org). However, for less-resourced languages like Icelandic, the situation is much less favourable. Icelandic language technology (LT) has really only existed for about a decade (Rögnvaldsson et al. 2009) and despite a rich lexicographic tradition there have until now been no specially LT-oriented semantic resources. Fortunately, over the last decade, the prerequisites for the application of (semi-)automatic methods in developing such semantic resources have now been created: a Part-of-Speech-tagger, a shallow parser and a lemmatizer (Loftsson 2008; Loftsson and Rögnvaldsson 2007; Ingason et al. 2008).

In 2007, a pilot study was run to extract semantic relations from an Icelandic dictionary (Nikulásdóttir and Whelpton 2009; Nikulásdóttir 2007a; Nikulásdóttir 2007b); following the success of this study and parallel developments in the field, a work-package for the creation of a database of semantic relations was in-

corporated into a major new project in Icelandic LT[1]. One central aim of the project is to experiment with known methods for the extraction of semantic relations and investigate how well they can be applied to Icelandic, given two significant factors: a) Icelandic is a highly inflected language; b) there are as yet no large corpora for the language. Most of the research in this area has focused on English which differs from Icelandic in both respects. To as great an extent as possible, we aim to exploit and develop methodologies which will be generally viable for other less-resourced languages with the support of open source tools.

The methods for the extraction of semantic relations or other semantic information can be divided into a) pattern-based methods and b) statistical methods. Pattern-based methods make use of syntactic and lexico-syntactic patterns as introduced by Hearst (1992), whereas statistical methods investigate statistical properties of language data. Following hybrid methodologies developed in recent years (Pantel and Pennachiotti 2008; Cimiano 2006; Cederberg and Widdows 2003) we intend to exploit and to combine methods from both approaches (Nikulásdóttir and Whelpton 2010).

In this paper we describe one statistical method for the extraction of semantic information, namely clustering on the basis of semantic relatedness. In sections 2 and 3 we discuss semantic relatedness and clustering in general. In section 4 we describe an experiment with the clustering of Icelandic nouns and how the results can be utilized for construction of a semantic database, as well as for lemma acquisition in traditional lexicography.

1 In 2009, the project *Viable Language Technology beyond English – Icelandic as a test case* received a three year Grant of Excellence from the Icelandic Research Fund *RANNÍS*

2. Semantic relatedness

Relations between words or concepts in semantic databases in the style of the Princeton WordNet (Fellbaum 1998) are predominantly classical semantic relations like hypernymy and synonymy. An extension of the set of classical relations is evolving in different resources: DanNet (Pedersen et al. 2009) e.g. uses the CONCERNS relation to express a general topic-relatedness of two concepts, (*goal* CONCERNS *sport*) and Boyd-Graber et al. (2006) have conducted experiments in enriching WordNet with a directed, weighted "evoking" relation. The evoking relation describes how strongly one concept evokes another one, e.g. *car* evokes *road*.

The Swedish SALDO resource (Borin and Forsberg 2009) is not designed along the lines of WordNet, but rather uses loosely characterized associative relations as its structuring principle.

We believe that a semantic database for NLP applications could profit from such loosely characterized relations alongside the classical semantic relations. One way to harvest such relations is to use measures of *semantic similarity* or *semantic relatedness*. The definition of semantic similarity has been rather vague (Manning and Schütze 1999:296), but it is now generally accepted that semantic similarity should be distinguished from semantic relatedness, which is a more general relation. Some scholars (cf. Resnik 1995:448) have treated the two kinds of relation as orthogonal to each other: so *car* and *wheel* are related by a specific classical relation – meronymy; and yet they are not similar to each other in the way that, say, *car* and *bicycle* are. We follow Budanitsky and Hirst (2001:29) in treating semantic relatedness as an umbrella term for a range of semantic relations including not only semantic similarity (*car~bicycle*) but also "lexical relationships such as meronymy (*car-wheel*) and antonymy (*hot-cold*), or [...] any kind of functional relationship or frequent association (*pencil-paper, penguin-*

Antarctica)." A more thorough discussion of semantic similarity and semantic relatedness can be found e.g. in Zesch and Gurevych (2009) and Turney (2006).

For the automatic extraction of semantic information, pattern-based methods (cf. Hearst 1992) are commonly used for the extraction of classical semantic relations such as hypernymy and meronymy. These methods use reliable lexico-syntactic patterns, like NP_1 *such as* NP_2 *((, NP)* (and* NP_n*))** and extraction rules. In this example a description of the respective rule would be "if NP_1 is followed by *such as* and NP_2 (and possibly an enumeration of nominal phrases), then NP_1 represents a hypernym of NP_2 (and the other NPs, if present). Given for example the sentence: *sports such as soccer, handball and basketball ...* one would extract (*soccer* HYPERNYM *sport*), (*handball* HYPERNYM *sport*) and (*basketball* HYPERNYM *sport*).

There are two main approaches to the automatic computation of the less well-defined semantic relatedness: a) to use knowledge sources like WordNet and Wikipedia, where paths between concepts or the glosses/definitions build the basis for measures of relatedness (Zesch and Gurevych 2009; Pedersen et al. 2004; Budanitsky and Hirst 2001; Resnik 1995); and b) to apply distributional methods to text corpora (Bullinaria 2008; Cimiano 2006; Weeds 2003; Cederberg and Widdows 2003; Manning and Schütze 1999). In this paper we follow the second of these approaches and describe an experiment using the distributional method of semantic clustering, based on co-occurrences of nouns and common content-bearing words (mainly nouns, verbs, adjectives). As is described in Section 3, the notion of "co-occurrence" used here is purely collocational (i.e. the co-occurrence of the target word with common content-bearing words); however, we also note the success of Cimiano (2006) and Weeds (2003) in measuring distributional similarity with respect to grammatical functions, especially with respect to direct objects. Application of this method for Icelandic awaits future work.

Such general semantic relatedness is important because it can be used as a "confidence measure" to validate specific semantic relations extracted with other methods (such as the pattern-based methods mentioned earlier). As an example, Cederberg and Widdows (2003) use general semantic relatedness to rank their hyponymy results, extracted using lexico-syntactic patterns. In doing so, they achieved a 30% reduction in error. We have already conducted initial experiments with this method, and intend to implement it on a broader basis (Nikulásdóttir and Whelpton 2010).

3. Distributional similarity and clustering

The fundamental assumption underlying distributional methods for computing semantic relatedness is that the semantic properties of a word determine the context they appear in. Thus, words appearing in a similar context are likely to be semantically related. The basic method therefore involves compiling distributional information on a set of *target words* with respect to some uniform definition of *context*.

3.1. The list of target words

The list of target words is generated using a text corpus. At the moment a balanced PoS-tagged, lemmatized corpus, *Mörkuð íslensk málheild* (MIM), is being developed at the Árni Magnússon Institute for Icelandic Studies (Helgadóttir 2004). The planned size of this corpus is about 25 million tokens, a reasonable size but still not especially large. For our present studies we use a subset of a preliminary version of this corpus (hereafter, SubMIM) containing about 8.8 million tokens, including punctuation marks etc. The source of this data is mainly newspaper texts (Morgunblaðið, a selection from the years 2000-2007), but further texts come from

a public science web portal at the University of Iceland (http://visindavefur.is), reports from Icelandic ministries, and from a medical Journal (Læknablaðið).

The tagging and lemmatizing was performed using the PoS-tagger *IceTagger* and the lemmatizer *Lemmald*, both included in the open source IceNLP-toolkit.[2]

The list of target words in our experiment is composed of those nouns which occur at least 18 times in SubMIM, approximately 11,500 nouns. To reduce noise, we removed the 100 most frequent nouns from the list, leaving 11,400 nouns. It should be noted that the original lists consisted of automatically lemmatized word forms. Incorrect lemmata were deleted but not corrected, and thus the lists do not mirror 100% the frequency relations in the corpus and the analysis does not account for wrongly lemmatized words. With the correction of the lemmatization these lists will change.

3.2. Defining context – words and windows

Semantic similarity is computed for a set of target words with respect to some uniform definition of context. There are a number of ways of defining "context". In our case, we assessed the co-occurrence of our target words with respect to a list of high frequency content-bearing words. Our initial plan was simply to use a general frequency list for Icelandic, purged of stop words[3] (1000 words total). However, this list includes words not present in the 2,000 most frequent words from SubMIM. It was therefore decided to replace those words which occurred on the general list but not on the SubMIM list with words from the SubMIM list, giving a hybrid 1,000 word list. To reduce noise, the 100 most frequent words were then removed, giving a final list of 900 words.

2 http://sourceforge.net/projects/icenlp

3 Another approach is taken by Bullinaria (2008), who does not remove stop words from the list of context words.

Information was then collected for every target noun on how often it occurred within a 25 word window of each of these 900 context words: i.e. the number of times a context word occurs within 12 words before or 12 words after a target noun. This information was represented in a matrix with the target nouns labeling the rows and the common content words labeling the columns. Each cell therefore contains the number of times a target noun (row) co-occurs with a content word (column). (See Table 1 below.)

Significant parameters which affect the characteristics of this co-occurrence matrix, and hence the overall results, include (cf. also Bullinaria 2008): the size and quality of the corpus being used and the information represented in the corpus, i.e. our corpus is part-of-speech tagged and lemmatized allowing us to extract the noun lemmata; the choice of the context words (the compilation of our list of 900 frequent content-bearing words represents in itself a range of significant choices); and the size of the co-occurrence window (e.g. co-occurrence could mean simple adjacency or a window of increasing length, up to perhaps 100 words).

To give a simple example of co-occurrence from our corpus, the target words *þvottahús* 'laundry room' and *baðherbergi* 'bathroom' co-occur with the context word *íbúð* 'flat, apartment' within a window of 25 words:

> ... hafa sérgeymslur inni í **íbúð**unum, bað með aðstöðu fyrir þvottavél, auk sameiginlegs þvottahúss og sameiginlegrar geymslu ...

> ... **íbúð**irnar afhendast fullbúnar án gólfefna en baðherbergi er flísalagt ...

In processing these snippets the program increments a counter for both target words in the cell representing the context word *íbúð*.

If *þvottahús* and *baðherbergi* generally share similar contexts in the corpus, i.e. both mostly appear near the same context words, their similarity value will be high[4]. Note how in the (fictive) co-occurrence matrix shown in Table 1 the top three target words share similar context, whereas *literature* and *cod* have different distributions:

	cw 1	*cw 2*	*cw 3*	*cw 4*	*cw 5*	*cw 6*
dining room	7	0	5	10	0	0
bathroom	11	0	9	9	0	0
laundry room	8	0	9	11	0	0
literature	0	8	0	0	0	0
cod	0	0	0	0	14	23

Table 1: An example of a co-occurrence matrix (cw = context word)

3.3. Similarity measures

To assess semantic relatedness, the rows in the co-occurrence matrix (representing the distribution of individual target nouns) must be compared. To increase the information content of the co-occurrence counts, the raw count data has to be transformed, e.g. by a logarithmic value (for different measures see e.g. Manning and Schütze 1999; Sahlgren 2006; Bullinaria 2008). The comparison of two rows follows through a *similarity measure.* The most common similarity measure used in measuring semantic relatedness is the so-called *cosine* similarity measure (cf. Manning and Schütze 1999). This measure gives a similarity value between 0 and 1, such that words with very different distribution have a similarity value closer to 0 but those with a very similar distribution have a similarity value closer to 1. Parameters which have an affect

4 These snippets also contain other target words, but for illustrative reasons only two are discussed in the example – likewise some words can be both context and target words, and in some approaches this holds for every word.

on results in this case concern the methods for transforming raw count data and the choice of similarity measure.

In summary, it is possible to process a corpus and to compute the similarity of the distribution of words of interest. Similarity values gained in this way are seen as an estimate of the semantic relatedness between words.

3.4. Clustering methods

The similarity information described so far concerns the similarity of individual word pairs, such as *laundry room* and *bathroom*. However, such information across all the nouns in a corpus can be used to produce clusters of nouns, where clustering is based on their relative similarity to each other. From a lexical semantic point of view, the ideal cluster would fall under a superordinate concept or semantic domain. An important challenge for clustering is that it is not known in advance how many such superordinate concepts or domains are evoked by nouns in the corpus.

A clustering algorithm must therefore group the words solely on the basis of their similarity values; identification and labeling of the superordinate domains which result from successful clustering must be performed manually. Continuing with the example of *laundry room* and *bathroom,* the algorithm might group these words together with *dining room, bedroom, child's bedroom, entrance, garage, wardrobe, parquet,* etc. It is, however, the task of a human assessor to label this cluster, for example with the concept HOUSE.

As with the other measures mentioned above, many clustering algorithms exist. In our experiment described in the next section we used the *k*-means clustering algorithm, with some adjustments. The interested reader can find good descriptions of *k*-means and other algorithms for example in Manning and Schütze (1999) and Duda et al. (2001).

4. Experiment: Clustering of nouns in Icelandic

In the following we describe an assessment of an experiment clustering Icelandic nouns according to semantic relatedness. We use two different statistics based on co-occurrence counts of words and cluster the data using adjusted *k*-means. First, an overview of the general approach and results is given, and then sections 4.2 and 4.3 describe manual assessment of selected clusters, an expert validation and a comparison to the *Icelandic Dictionary* respectively.

4.1. General results

We use two different measures to transform the raw co-occurrence counts in the co-occurrence matrix containing rows of target words and columns of context words: a logarithmic measure (Manning and Schütze 1999:302) and Positive Pointwise Mutual Information, PPMI (Bullinaria and Levy 2007)[5]. As a result two distinct matrices are produced, a) the *log* co-occurrence matrix and b) the PPMI co-occurrence matrix. These matrices form the input for the clustering algorithm, where the cosine similarity measure is used to compute similarity of distributions (i.e. rows in the matrix). The results of the clustering based on the *log* co-occurrence matrix show 79 clusters, each containing from 9 to 192 words, whereas the PPMI matrix results in 76 clusters with 4 to 198 words. The cluster content is automatically ranked according to how close a word is to the word considered being "in the middle" of the cluster.

5 PPMI is a statistical method based on conditional and overall probabilities, used to increase the information content of raw co-occurrence data. We avoid technical details here but interested readers can find a more thorough account in the reference provided.

From a shallow screening of the clusters, it seems that 60 clusters from the *log* matrix can be characterized by a subsuming concept and 59 clusters from the PPMI matrix. As the clustering is completely unguided – i.e. no pre-defined categories are given – the resulting clusters have different ontological status: scientific domains (BIOCHEMISTRY, BIOLOGY), concrete things (HOUSE, VEHICLE), domains containing mostly proper nouns (FOOTBALLERS, MUSIC/MUSICIANS) domains from public discourse (POLITICS, FINANCES/BUSINESS), etc.

We will now report on a preliminary manual assessment of these clusters. Lacking a gold standard to evaluate these results, we selected one domain for manual assessment, the domain of finances and business. In the following we describe the examination of these clusters.

4.2. Expert validation

For the expert validation, two employees of a bank were asked to rate relatedness of words to the domain of finances and business.

4.2.1. Clusters for expert validation

The clusters used in this part of the assessment come from the partition based on the *log* transformation measure. This partition includes five clusters related to the domain of finances and business, each of them containing from 74 up to 173 words, all in all 555 words. We selected the 50 top words from each cluster for the validation.

4.2.2. Validation by finance professionals

To evaluate to what extent the words in the clusters are related to the domain of finance and business, two employees from a bank were asked to rate the words from the clusters described in *Section*

4.2.1[6]. Four scores were possible: 3 – very related to the domain; 2 – fairly related to the domain; 1 – not particularly related to the domain; 0 – not at all related to the domain. This method produced an interesting result with respect to domain-specificity: the bankers were both so focused on the banking domain that words obviously related to the business domain, like *samsteypa* 'conglomerate' and *sölufyrirtæki* 'a selling company', were rated as 0. Even company names (such as the well-known bank, *Glitnir*) were mostly rated with 0, sometimes with 1. Given this domain-bias, only the results from the two clusters directly related to banking proved to be useful and they are the focus of this assessment.

The inter-annotator agreement with respect to the four scoring categories was 65%; however, grouping the positive scores (very/fairly related) and the negative scores (little/not-at-all related) increased inter-annotator agreement to 80%.

score	banker 1	banker 2
3	58	55
2	22	16
1	16	25
0	4	4

Table 2: Rating results from two employees of a bank, rating words in two clusters, each containing 50 words.

As shown in Table 2, according to *banker 1,* 80% of the words in the two clusters are very or fairly related to the banking domain and 71% according to *banker 2*. Taking only the inter-annotator agreement into account, 64% of the words in the two clusters are very or rather related to the domain. This degree of agreement is not especially high, though one has to bear in mind, that "there exist numerous equally valid alternative ways" of doing manual categorization (Bullinaria 2008:5).

6 We want to thank the two employees at *Íslandsbanki* for their assessment.

4.3. Comparison with the Icelandic Dictionary (ÍO)

The purpose of the comparison with ÍO is twofold: to find out a) whether a data-driven method like this adds anything to the information already in the dictionary and b) whether the classification of those words present in the dictionary matches the clustering results[7]. This part of the assessment also concerns the domain of finance and business.

4.3.1. The Icelandic Dictionary (ÍO)

ÍO originates in the first general monolingual Icelandic dictionary from 1963: *Íslenzk orðabók handa skólum og almenningi* 'An Icelandic Dictionary for Schools and the General Public'. All later editions and revisions build on this version and no general modernisation has taken place.

The following assessment concerns the coverage of lemmata in ÍO. We will therefore report briefly on the lemma list and definition vocabulary. A substantial part of the lemmata in the first version (1963) comes from a bilingual Icelandic-Danish dictionary from 1920-1924: *Íslensk-dönsk orðabók* (Kvaran 1998). Even in the second edition of ÍO (1983), there is a bias towards dated vocabulary, but we are not aware of whether the planned correction of this bias back towards contemporary usage (Árnason 1998) has been performed. Árnason also notes that ÍO has inconsistent definition texts and unclear objectives for the selection of lemmata. It has furthermore been determined that c. 42% of the definition vocabulary did not form lemmata in the 1983 edition (Bjarnadóttir 1998). There is a lexicographical rule stating that derivatives and compounds built by productive word formation rules where meaning and form are predictable from the parts need not become

7 We started the comparison also using *Stóra orðabókin um íslenska málnotkun* 'The Large Dictionary of Icelandic Language Use' (Jón Hilmar Jónsson 2005), but the coverage of the finance/business domain words was too small to be of use.

lemmata in a standard dictionary. However, this rule has not been consistently followed in ÍO (cf. Bjarnadóttir 1998:39):

> **eiturbikar** bikar með eitri í ***poison goblet*** *a goblet containing poison*
> **eplakaka** kaka með eplum í ***apple cake*** *a cake containing apples*

Yet even taking into account this rule (and other lexicographical lemma selection rules), Bjarnadóttir concludes that about 6,300 lexemes, or 12% of the definition vocabulary, are missing from the lemma list. It is thus apparent that the construction of the lemma list in ÍO has not followed strict guidelines, and the reasons for a word being included or not included in the lemma list are not always clear-cut.

Since the year 2000, ÍO has been accessible on the web (http://snara.is) and that is the version which we use for our experiment.

4.3.2. Clusters for the dictionary comparison

Both partitions (the one based on *log* and the one based on PPMI) include five clusters related to the domain of finances and business. The *log*-clusters have a total of 555 words and the PPMI-clusters a total of 582 words, with 458 words being common to both partitions. As with the clusters for the expert validation, the 50 top words from each cluster were selected. The resulting lists were merged, thus erasing the partition into distinct clusters and removing duplicates, and all proper nouns were deleted. The final list for the comparison with ÍO consists of 260 common nouns.

4.3.3. Lemma coverage

The first test concerns the question of how many words from the test list are lemmata in ÍO. Of the 260 selected words, 147 or 56.5% appear as a lemma in ÍO. Another 17 words, all compounds, are included in one of the compound lists within the definition of many lemmata. This means that 96 words or 36.9% are not listed

in the dictionary. Some of these words are productive compounds, often avoided in dictionaries, like compounds with *heildar-* 'total': *heildarvelta* 'total turnover', and *heildarútflutningur* 'total export' (but recall that the dictionary lemma list also contains many productive compounds, see section 4.3.1). Given that our resource is intended for NLP applications, we want our database to include such compounds, as this will save applications from having to decode them individually.

Another possible reason for a word being or not being in the dictionary is frequency. The 260 words in the finance/business list appear from 18 to 134 times in SubMIM. The ten most frequent words are all lemmata in the dictionary, but other high frequency words like *markaðsvirði* 'market value' (81 occurrences), *hönnunarfyrirtæki* 'design company' (80 occurrences), *eignarhaldsfélag* 'holding company' (70 occurrences) and *yfirtökutilboð* 'take-over bid' (70 occurrences) are not. Also low frequency words are either in the dictionary (*smásöluverð* 'retail price') or not (*smásölustig* 'retail level') – both words appear only 18 times in the corpus.

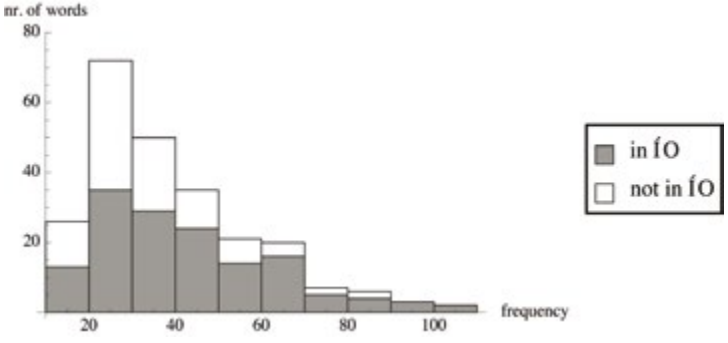


Figure 1: Frequencies of words from the finance/business domain listed and not listed in ÍO (the highest frequency numbers are left out to make the graph better readable)

Figure 1 shows the frequency distribution of the 260 words, separated by occurrence and non-occurrence in ÍO. This suggests

that our clustering technique does indeed provide useful results for extending dictionary coverage.

4.3.4. Domain labeling

The second evaluation task regarding ÍO was to compare the classification of the words listed in the dictionary with the cluster domain. ÍO has the domain label *viðskipti/hagfræði* 'business/economics'. Of the 147 words listed as lemmata, 52 or 35.4% have this domain assignment. Four words are assigned to the domain *stjórnsýsla* 'administration'; *ál* 'aluminium' and *dísilolía* 'diesel' are assigned to *eðlis-/efnafræði* 'physics/chemistry', though they can also be seen as related to economics. Three unrelated words have other domain assignments. That leaves 86 words without any domain assignment: of these, words like *lánsfé* 'loan capital', *afborgun* 'amortization', and *bankareikningur* 'bank account'are all strongly related to the finance domain; however, these 86 words also include items that are completely unrelated to the finance domain, such as *kindakjöt* 'mutton', *vín* 'wine', and *varahlutur* 'spare part'. Once again, our clustering technique does provide useful results for extending the domain information in the dictionary, though for this very reason it makes the dictionary an ineffective reference point for the assessment of cluster quality.

5. Conclusion

We have described one method for extracting semantic information from text. These results will contribute to the development of a semantic database for Icelandic language technology. The methods described here will be used alongside other techniques (cf. Nikulásdóttir and Whelpton 2010), in the belief that a hybrid methodology (Pantel and Pennachiotti 2008; Cimiano 2006; Cederberg and Widdows 2003) will yield the highest quality results

from limited resources. The results reported here also suggest that clustering by semantic relatedness can be of great use to traditional lexicography in discovering potential lemma candidates.

References

Árnason, Mörður 1998: Endurútgáfa "Íslenskrar orðabókar". Stefna – staða – horfur. In: *Orð og tunga* 4:1-8.

Bjarnadóttir, Kristín 1998: Um skýringarorðaforðann. In: *Orð og tunga* 4:33-44.

Blöndal, Sigfús 1920-1924: *Íslensk-dönsk orðabók.* Reykjavík.

Borin, Lars and Markus Forsberg 2009: All in the Family: A Comparison of SALDO and WordNet. In: Bolette Sandford Pedersen et al. (eds): *Proceedings of the NODALIDA 2009 Workshop Wordnets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies,* vol. 7 of *NEALT Proceedings Series,* Odense, Denmark, 7-12.

Boyd-Graber, Jordan; Christiane Fellbaum, Daniel Osherson, and Ropert Schapire 2006: Adding Dense, Weighted Connections to WordNet. In: Petr Sojka et al. (eds): *Proceedings of the GWC,* 29-35.

Budanitsky, Alexander and Graeme Hirst 2001: Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of five Measures. In: *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001),* Pittsburgh, PA, 29-34.

Bullinaria, John A. 2008: Semantic Categorization Using Simple Word Co-occurrence Statistics. In: M. Baroni et al. (eds): *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics,* Hamburg: ESSLI, 1-8.

Bullinaria, John A. and Joseph P. Levy 2007: Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. In: *Behavior Research Methods,* 39:510-526.

Cederberg, Scott and Dominic Widdows 2003: Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction. In: *Proceedings of the International Conference on Natural Language Learning (CoNLL)*, 111-118.

Cimiano, Philipp 2006: *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.* Springer.

Duda, Richard O.; Peter E. Hart and David G. Stork 2001: *Pattern Classification.* New York, Chichester etc.: John Wiley.

Fellbaum, Christiane (ed.) 1998: *WordNet. An Electronic Lexical Database.* Cambridge Mass., London: MIT Press.

Hearst, Marti A. 1992: Automatic Acquisition of Hyponyms from Large Text Corpora. In: *Proceedings of COLING-92*, Nantes, 539-545.

Helgadóttir, Sigrún 2004: Mörkuð íslensk málheild. In: *Samspil tungu og tækni.* Reykjavík: Ministry of Education, Science and Culture, 65-71.

Íslenzk orðabók handa skólum og almenningi 1963. Árni Böðvarsson (ed). Reykjavík: Menningarsjóður.

Íslensk orðabók handa skólum og almenningi 1983. 2nd ed. Árni Böðvarsson (ed). Reykjavík: Menningarsjóður.

Ingason, Anton Karl; Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson 2008: A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In: Bengt Nordström et al. (eds): *Advances in Natural Language Processing,* vol. 5221 of *Lecture Notes in Computer Science,* Berlin: Springer, 205-216.

Kvaran, Guðrún 1998: Uppruni orðaforðans í "Íslenskri orðabók". In: *Orð og tunga* 4:9-16.

Loftsson, Hrafn 2008: Tagging Icelandic Text: A Linguistic Rule-Based Approach. *Nordic Journal of Linguistics,* 31(1):47-72.

Loftsson, Hrafn and Eiríkur Rögnvaldsson 2007: IceParser: An Incremental Finite-State Parser for Icelandic. In: Joakim Nivre et al. (eds): *Proceedings of the 16th Nordic Conference on Computational Linguistics (NODALIDA),* 128-135.

Manning, Christopher and Hinrich Schütze 1999: *Foundations of Statistical Natural Language Processing.* Cambridge Mass., London: MIT Press.

Nikulásdóttir, Anna Björk and Matthew Whelpton 2010: Extraction of Semantic Relations as a Basis for a Future Semantic Database for Icelandic. In: *Proceedings of the 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, Malta.

Nikulásdóttir, Anna Björk and Matthew Whelpton 2009: Automatic Extraction of Semantic Relations for Less-Resourced Languages. In: Bolette Sandford Pedersen et al. (eds): *Proceedings of the NODALIDA 2009 Workshop Wordnets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies,* vol. 7 of *NEALT Proceedings Series,* Odense, Denmark, 1-6.

Nikulásdóttir, Anna Björk 2007a: *Automatische Extrahierung von semantischen Relationen aus einem einsprachigen isländischen Wörterbuch* [Automatic Extraction of Semantic Relations from a Monolingual Icelandic Dictionary], Master's thesis, University of Heidelberg.

Nikulásdóttir, Anna Björk 2007b: Sjálfvirk greining merkingarvensla í *Íslenskri orðabók*. In: *Orð og tunga* 9:5-24.

Pantel, Patrick and Marco Pennacchiotti 2008: Automatically Harvesting and Ontologizing Semantic Relations. In: Paul Buitelaar and Philipp Cimiano (eds): *Ontology Learning and Population: Bridging the Gap between Text and Knowledge – Selected Contributions to Ontology Learning from Text.* IOS Press.

Pedersen, Bolette S.; Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen 2009: DanNet: the Challenge of Compiling a Wordnet for Danish by Reusing a Monolingual Dictionary. In: *Language Resources & Evaluation*, 43:269-299.

Pedersen, Ted; Siddharth Patwardhan, and Jason Michelizzi 2004: WordNet::Similarity – Measuring the Relatedness of Concepts. In: *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI-04),* San Jose, CA, 1024-1025.

Resnik, Philip 1995: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: *Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95),* San Mateo, CA. San Francisco: Morgan Kaufmann, 448-453.

Rögnvaldsson, Eiríkur; Hrafn Loftsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Anna Björk Nikulásdóttir, Matthew Whelpton, and Anton Karl Ingason 2009: Icelandic Language Resources and Technology: Status and Prospects. In: Rickard Domeij et al. (eds): *Proceedings of the NODALIDA 2009 Workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources,* vol. 5 of *NEALT Proceedings Series*, Odense, Denmark, 27-32.

Sahlgren, Magnus 2006: The Word-Space Model. Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High Dimensional Vector Spaces. PhD thesis, Stockholm University.

Turney, Peter 2006: Similarity of Semantic Relations. In: *Computational Linguistics* 32(3), 379-416.

Weeds, Julie 2003: *Measures and Applications of Lexical Distributional Similarity.* Ph.D. thesis, University of Sussex.

Zesch, Torsten and Iryna Gurevych 2009: Wisdom of Crowds versus Wisdom of Linguistics – Measuring the Semantic Relatedness of Words. In: *Natural Language Engineering,* 16(1):25-59.

Internet references

ÍO = *Íslensk orðabók* (2007). Mörður Árnason (ed). 4th edition. Reykjavík: Edda http://snara.is (March 2010).

Jón Hilmar Jónsson (2005): *Stóra orðabókin um íslenska málnotkun.* Reykjavík: JPV. http://snara.is (March 2010).

Anna Björk Nikulásdóttir
Ph.D. student
University of Iceland
Sæmundargötu 2
IS-101 Reykjavík
abn@hi.is

Matthew Whelpton
Associate Professor
University of Iceland
Sæmundargötu 2
IS-101 Reykjavík
whelpton@hi.is

Semantiske sprogressourcer – mellem sprogteknologi og leksikografi

Bolette Sandford Pedersen

This paper discusses the synergy between lexicography and semantic language resources meant for computational use; before, now and in the future. On the basis of a brief historical overview of the backgrounds for language technology and lexicography, respectively, I analyze why the two fields have not always cooperated as closely as one would think useful. I give a recent example of a project that has exploited the similarities between the two fields by reusing a monolingual dictionary for the compilation of a Danish wordnet for technological use: DanNet. I describe some areas where modifications have been necessary in the reuse process; this regards in particular the adjustment of hyponymy hierarchies and the spelling-out of underspecified information. I conclude that the two fields will most presumably be much more connected in the future due to recent corpus and editing tools which help exploit more radically the intersection between the two areas.

1. Indledning

I udgangspunktet må det være helt naturligt at sprogteknologien straks vender sig mod leksikografien når der opstår behov for leksikografiske data i større mængder. Selvom almindelige ordbøger er beregnet til mennesker mens maskinanvendelige ordbaser har computeren som mellemled i form af et program der udnytter dem, må mængden af fællesviden være stor. I denne artikel fokuserer jeg på semantisk viden og på i hvor høj grad der er synergi mellem den semantiske viden vi udtrykker ved hjælp af definitioner og brugseksempler i menneskeordbøger, og den semantik vi udtrykker i maskinanvendelige ordbaser i form af unifikationsba-

serede netværk. Mit udgangspunkt for denne redegørelse er DanNet, det leksikalsk-semantiske wordnet som vi på Københavns Universitet og Det Danske Sprog- og Litteraturselskab har udviklet i årene 2005-2010 og som kan downloades som open source fra wordnet.dk (se også Pedersen et al. 2009)[1].

I udviklingen af denne leksikalske ressource har vi i høj grad forsøgt at udnytte den formodede fællesmængde af viden i de to typer ressourcer idet vi har tilstræbt genanvendelse af definitioner og oplysninger om nærmeste danske overbegreb fra Den Danske Ordbog (DDO), som er en moderne, korpusbaseret ordbog for dansk.

Artiklen indledes med et historisk tilbageblik på sprogteknologiens udspring i det formelle sprogsyn med det formål at afdække hvorfor synergien mellem leksikografi og sprogteknologi ikke altid har været åbenbar. Jeg refererer bl.a. til en artikel af Ide & Véronis (1995) som har fået en vis betydning for synet på genanvendelse af leksikografiske data til sprogteknologiske formål, og jeg ser bl.a på Pustejovskys generative leksikon som et moderne bud på hvordan leksikografiske data behandles i den formelle sprogtradition.

Dernæst beskrives DanNets tilblivelse med udgangspunkt i DDO. Der fokuseres på to typer af justeringer som har været nødvendige for at sikre at den semantiske viden fra DDO får en passende form i et formelt baseret wordnet: Tilpasning af inkonsistente eller underspecificerede hyponymier samt tilføjelse eller eksplicitering af manglende information. Endelig beskrives i sidste afsnit nogle anvendelsesperspektiver for semantiske sprogressourcer.

1 DanNet-projektet er støttet af Statens Humanistiske Forskningsråd, og ressourcen videreudvikles nu under DK-CLARIN-projektet støttet af Forskningsrådet for Kultur og Kommunikation.

2. Sprogteknologiens udspring i det formelle sprogsyn

I modsætning til leksikografien, som udspringer af et beskrivende sprogsyn forankret i en humanistisk tradition, så udspringer datalingvistik/sprogteknologi og den deraf afledte datamatiske leksikografi i højere grad af det såkaldt formelle sprogsyn. Et karakteristikum for det formelle sprogsyn er at man her anser sproget som et produktionssystem som fx udtrykt i Chomskys generative grammatikker. Disse systemer har nær tilknytning til matematikkens og logikkens udvikling af formelle, logiske systemer, og med den hastige udvikling af computeren i sidste halvdel af forrige århundrede blev denne et håndgribeligt udtryk for hvordan man kan formalisere sprog og tænkning. Man kan altså sige at udgangspunktet her i højere grad er naturvidenskabeligt eller i hvert fald placerer sig i et grænseområde mellem datalogi og naturvidenskab på den ene side og humaniora på den anden (jf. bl.a. Prebensen 2006 og Spang-Hanssen 2006). Traditionelt har (datamatisk) leksikografi ikke spillet nogen stor rolle i de formelle teorier; først med de leksikalistiske teoriers fremkomst i firserne (Kaplan & Bresnan 1982, Pollard & Sag 1989) kom ordforrådet og dets informationer i fokus. Trenden gik nu fra at betragte ordbogen som et mere eller mindre uinteressant appendiks til den formelle grammatik til at indeholde langt mere strukturel information om fx valens, argumentstruktur og senere selektionsrestriktioner.

2.1. Semantikken i det formelle sprogsyn

Hvis vi ser mere specifikt på semantikken i den formelle sprogtradition, bliver det også klart at interesseområderne inden for dette felt relaterer sig stærkt til logikken og de genstande som typisk behandles her. Cann (1993) udtrykker fx at det at kende betydningen

af en ytring er at forstå dens sandhedsforudsætninger. Ytringers sandhedsværdi er et centralt parameter i formel semantik sammen med aspekter som *referens* (fx pronominel referens), *virkefelt* (fx negationers virkefelt) og *kausalitet* (fx mellem hoved- og bisætning når de forbindes med forskellige konjunktioner). Interesseobjekter er med andre ord typisk de elementer i sproget der har en parallel til logikkens elementer; det vil primært sige funktionsordene, så som determinatorer og ubestemte pronominer der minder om kvantorer i logikken, samt konjunktioner (*og, eller, hvis, så*) og visse adverbier. Indholdsordene, derimod, som er leksikografiens hovedfelt, altså især substantiver, verber og adjektiver, har derimod været stedmoderligt behandlet i den formelle semantik; oftest blot repræsenteret ved hjælp af det underspecificerede semantiske mærke (') placeret efter et indholdsord som reference til dets indhold[2].

Vel hjulpet af Partee som i høj grad dannede bro imellem formel og leksikalsk semantik, indledte Boguraev & Briscoe (1989), Copestake & Briscoe (1995), Calzolari (1988) m.fl. i firserne og begyndelsen af halvfemserne en ny æra i og med at de begyndte at opstille formelle, unifikationsbaserede teorier for en leksikalsk semantik der havde indholdsordene i centrum og kunne interagere med andre teorier inden for formel syntaks. AQUILEX-projektet (jf. fx Copestake 1990) var et EU-forskningsprojekt der fik stor betydning i denne sammenhæng idet det var et af de første projekter som samtidig med teoriudvikling reelt påbegyndte udviklingen af større leksikalske ressourcer med semantisk information til maskinel anvendelse.

Pustejovskys formelle beskrivelsesapparat til håndtering af leksikalsk semantik (1995) fik også stor gennemslagskraft. Det

2 Fx som hos Cann (1993:38): *poison' (ethel', the_cat')* hvor der primært fokuseres på relationen mellem prædikat og argumenter mens verbets og substantivernes øvrige betydningskomponenter underspecificeres.

blev udarbejdet med henblik på dels at kunne håndtere forskellige betydningsdimensioner og opløse flertydighed, dels at kunne beregne semantisk nærhed og afhængighedsrelationer imellem begreber. Pustejovsky omtaler ordenes kernesemantik som "qualiastruktur": en struktur der indbefatter dimensioner som *form*, *oprindelse* og *formål*. Denne kernesemantik menes at interagere med en række generative mekanismer som således forklarer hvordan begreber tager farve af hinanden i kontekst. Når vi fx omtaler en *bil* som *hurtig*; refererer vi til dens køreegenskaber, mens når vi omtaler en *hurtig læser* refererer til læsehastighed. Dette fænomen omtales som selektiv binding; egenskaben *hurtig* "bindes til" kernesemantikken i det substantiv det lægger sig til som for substantiverne *bil* og *læser* indbefatter hhv. *køre* og *læse*.

Pustejovskys teorier dannede udgangspunkt for det lidt senere EU-projekt SIMPLE, som implementerede centrale dele af qualiastrukturen (Lenci et al. 2000). SIMPLE havde en dansk (se Pedersen & Paggio 2004) og en svensk aflægger (Kokkinakis et al. 2000) og senere en norsk, og den meget omfattende semantiske struktur i dette projekt dannede i høj grad forbillede for DanNet, som indbefatter aspekter fra SIMPLE der ikke indgår i traditionelle wordnets (bl.a. elementer fra qualiastrukturen).

2.2. Den leksikalske semantik og det mentale leksikon

Nogenlunde samtidig med denne udvikling inden for det formelle sprogparadigme påbegyndte Miller og Fellbaum (Fellbaum 1998) deres psykolingvistiske eksperimenter omkring det mentale leksikon i form af wordnets. Med udgangspunkt i to sprogpsykologiske hypoteser, nemlig adskillelseshypotesen og mønsterhypotesen, påbegyndte man udviklingen af det såkaldte Princeton WordNet, som er et netværk af begreber lagret i en databasestruktur. Adskillelseshypotesen går ud fra en antagelse om at det mentale leksikon har selvstændig status i forhold til fx vores evne til at producere

grammatisk korrekte sætninger, og at det derfor giver mening at studere ordforrådet, altså det mentale leksikon, uafhængigt af vores øvrige sprogevne, fx vores grammatiske. Mønsterhypotesen bygger på antagelsen om at menneskers viden om ords betydninger er lagret i form af netværk med indbyrdes semantiske relationer imellem begreber. Wordnets med sådanne semantiske relationer – eller leksikalsk-semantiske net som de også kaldes – er efterfølgende blevet en stor succes inden for især sprogteknologien og udvikles nu for en lang række sprog inklusive flere af de nordiske.

2.3. Hvor kommer leksikografien ind?

Som allerede nævnt kan man undre sig over at almindelige menneskeordbøger og den semantik de indeholder, ikke fra starten er blevet anvendt langt mere når der skulle bygges datamatiske ordbaser og leksikalske net til datamatisk anvendelse. Ordbaserne har i lang tid udgjort en flaskehals i sprogteknologiske applikationer bl.a. fordi deres dækningsgrad langtfra har været god nok, og hvad var mere naturligt end at genbruge fra andre, store ordressourcer?

Selvom Boguraev & Briscoe (1989) allerede i slutningen af firserne beskriver muligheder for genbrug af almindelige ordbøger, og selvom genbrugsaspektet indgår i den første fase af AQUILEX-projektet, så evalueres genbrugsaspektet i 1995 negativt af Ide & Véronis. Deres vurdering er relativt nedslående, og denne evaluering har muligvis påvirket udviklingen negativt da der er tale om et af de eneste bredere studier af området fra den periode. Forfatterne konkluderer bl.a. at den information man kan udlede fra ordbøger, er for inkonsistent og usystematisk til at det kan betale sig at forsøge at udtrække den[3]. Udtrækning af hierarkier og se-

3 Det er værd at bemærke at selvom ordbogsproduktion generelt har været inde i en rivende udvikling bl.a. på grund af moderne editerings- og korpusværktøjer, og ordbøgerne således er blevet langt mere konsistente

mantiske relationer giver altså et for ujævnt resultat som kræver for meget efterredigering efter deres skøn, og de foreslår derfor at man i stedet anvender viden fra korpora, som da også var baggrundsressourcen for den anden fase af Aquilex-projektet. Meget semantisk information er også underforstået i ordbøger fordi man regner med sprogbrugerens forhåndsviden, og den eksplicitering som er nødvendig for computeranvendelse er blevet anset for at være for omfattende til at genbrug kunne betale sig.

En anden, relateret diskussion går på i hvor høj grad de betydningsdistinktioner der ses i almindelige ordbøger, kan genanvendes til maskinelle formål, som beskrevet i Kilgarriff (1997) og senere i Ide & Wilks (2007). Er de for finkornede? Er de korpusbaserede i tilstrækkelig grad? Bør betydningsdistinktioner beskrives på en helt anden måde end ved at opsplitte i betydninger og underbetydninger? Diskussionerne er mange, og også i wordnetkredse er der en livlig debat om hvorvidt det kan betale sig at tage udgangspunkt i almindelige, monolingvale ordbøger når man skal udvikle wordnets for nye sprog, eller om man hellere skal oversætte Princeton WordNet til de pågældende sprog for derefter at efterjustere monolingvalt. På NFL's årlige symposium på Schæffergården i 2010 fremgik det at der i øjeblikket udvikles wordnets for de nordiske sprog med begge metoder. I det følgende afsnit beskriver jeg fordele og ulemper ved anvendelse af den monolingvale metode, som er den vi har valgt for udviklingen af DanNet.

3. Genbrug af en monolingval ordbog i sprogteknologisk sammenhæng

Selvom det har været en stor hjælp og altså efter vores bedste overbevisning besværet værd at genbruge en eksisterende ordbog, skal

end tidligere, så vil der være tale om en vis forsinkelse inden dette slår igennem inden for sprogteknologien.

der ikke lægges skjul på at der har været behov for en del tilpasninger undervejs i stil med dem som andre allerede har erfaret, og som er kort skitseret ovenfor. I det følgende vil jeg opridse nogle af de vigtigste justeringer og kompletteringer.

3.1. Tilpasning af inkonsistente eller underspecificerede hierarkier

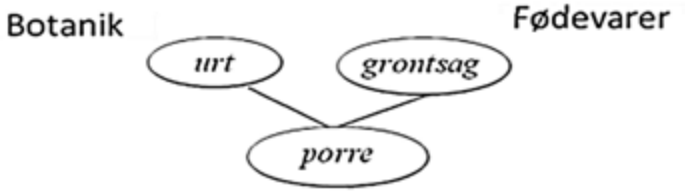
Det vigtigste genbrugsaspekt i udviklingen af DanNet har været brug af genus proximum-oplysninger i DDO til skabelsen af den overordnede struktur i det semantiske net. Hvis vi fx ser på et semantisk område som frugt og grønt vil man se at udgangspunktet i DDO er noget forskelligartet idet der for denne gruppe er fire forskellige beskrivelsesstrategier (jf. Pedersen, Nimb & Braasch 2010):

1. Lemma med kun én betydning i DDO; overbegreb er grøntsag eller rodfrugt: Fx *avokado, majroe*
2. Lemma med kun én betydning i DDO; overbegreb er plante: Fx *artiskok, spinat*
3. Lemma med to betydninger i DDO, med hhv. plante og grøntsag som overbegreb. Fx: *tomat*, *græskar*
4. Lemma med over- og underbetydning i DDO, med hhv. plante og plantedel som overbegreb: Fx: *gulerod, jordskok*

Det man ser af denne forskelligartede måde at beskrive grøntsager i DDO på er at man – ud over at vakle mellem overhovedet at beskrive en madbetydning som en selvstændig betydning – vakler imellem hvorvidt man skal beskrive fødevarebetydningen af en plante eller plantedel ud fra en *fødevarekategorisering* som er tilfældet i 1 og 3 (*grøntsag, rodfrugt* som overbegreb), eller man skal anvende den *botaniske taksonomi*, som det er tilfældet i 4 (*plantedel*). For at få en bedre struktur på dette i netværket forsøger vi i Dan-

Net at etablere to parallelle taksonomier: en botanisk taksonomi og en fødevaretaksonomi. Den første kalder vi i henhold til Cruse (2002) for en naturlig (biologisk) taksonomi, mens fødevaretaksonomien er en funktionstaksonomi hvor genstandens anvendelse (i dette tilfælde som fødevare) er det strukturerende princip[4].

Overbegreber der kun indgår i fødevaretaksonomien, er begreber som *grøntsag, suppeurt,* og *krydderurt*, mens begreber som fx *rod, stenfrugt* og *skærmplante* kun figurerer i den botaniske taksonomi. I et forsøg på at opnå en mere systematisk struktur i DanNet end den man finder i DDO, bestræber vi os på at repræsentere de planter eller plantedele som vi typisk spiser i vores kultur i *begge* taksonomier. Således *porre* som i DanNet dels har overbegrebet



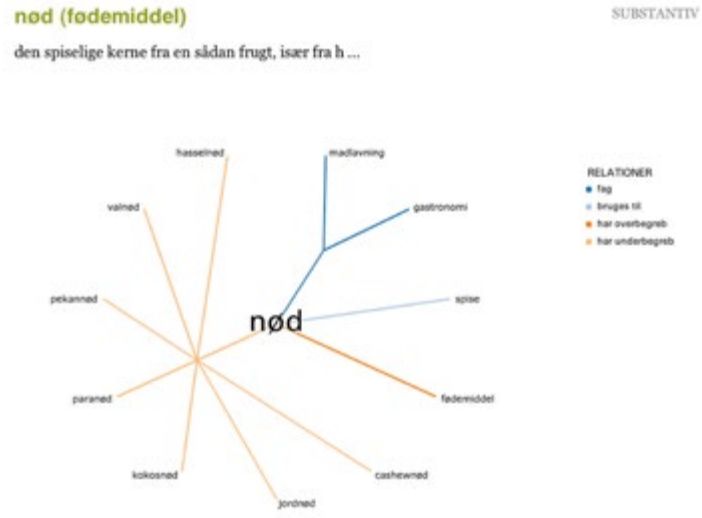
grøntsag/grønt fra fødevaretaksonomien, dels overbegrebet *urt* fra den botaniske taksonomi; se figur 1:

Figur 1: *porre* nedarver dels fra den botaniske taksonomi, dels fra fødevaretaksonomien

For at gøre forvirringen komplet findes der også en række udtryk som refererer til forskellige begreber i de to taksonomier, og som viser at lægmand og fagmand (fx botanikeren) har forskellige strukturelle opfattelser af domænet. *Bær, nødder* og *frugter* udpeger fx forskellige genstande hos hhv. en botaniker og en lægmand. For en botaniker er en *tomat* således en *frugt* og et *jordbær* en *nød* (som igen er en *frugt*), hvilket går imod den almindelige opfattelse

4 Wierzbicka (1996) har en lignende skelnen mellem *naturlige* og *kulturelle* typer.

hvor vi anser en *tomat* for at være en *grøntsag* og et *jordbær* for et *bær*. "Falske venner" som *bær*, *nød* og *frugt* holdes adskilt ved at etablere én betydning i hver taksonomi; dog skal det bemærkes at DanNet langt fra afbilder en komplet biologisk taksonomi på disse områder da resursen generelt behandler almensproget og derfor fx



beskriver fødevarer mere uddybende. Figur 2 angiver *nød* i fødevaretaksonomien.

Figur 2: Fødevarebetydningen af *nød* i DanNet

En mere uddybet redegørelse for hvordan fødevarer, botanik og zoologi er behandlet i DanNet gives i Pedersen, Nimb & Braasch (2010).

3.2. Reorganisering af synonymibegrebet – en pragmatisk tilgang

DanNet har også et andet og mere pragmatisk synonymibegreb end DDO. Det er typisk for wordnets at synonymer tolkes noget

bredere end i traditionelle ordbøger. I DanNet har vi i høj grad set på paralleliteten i søsterstrukturen som den forekommer i DDO. Når begreberne *fag*, *videnskab*, og *lære* i DDO har parallelle "søstre", nemlig hhv. *informatik*, *bromatologi* under *lære*, *samfundsfag* under *fag* og *datalogi* som undertype til *videnskab*, så er det en indikation om at overbegreberne bruges mere eller mindre i flæng. Fra et praktisk, sprogteknologisk synspunkt er det mest nærliggende at foretage en sammenlægning i ét såkaldt "synset" (= mængde af synonymer der betegner samme begreb): {*lære, fag, videnskab*} og have de forskellige fag som søstre under samme overbegreb. *Anordning* og *indretning* er andre eksempler på overbegreber i DDO som parallelle søsterbegreber refererer til som nærmeste overbegreb, fx hhv. *støttefod* og *lampeskærm*.

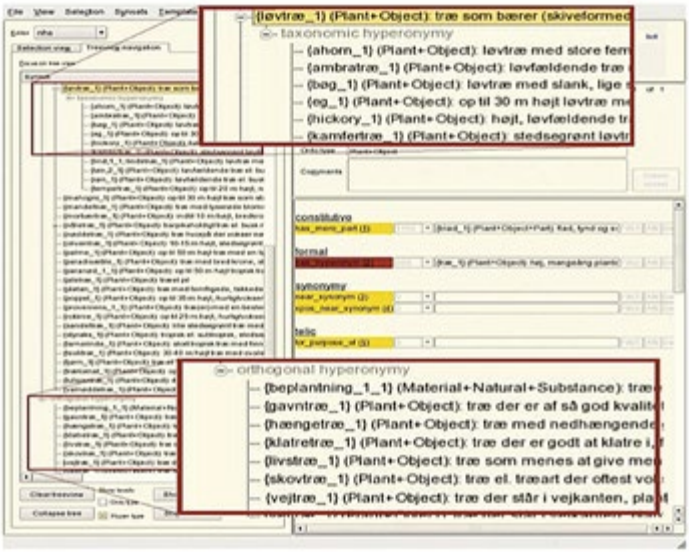
I praksis er disse typer af synonymi imidlertid ikke altid lige lette at identificere. Derfor vil man stadig finde begreber i DanNet som set fra en sprogteknologisk synsvinkel med fordel kunne sammenlægges. Dette er en af de opgaver som vi håber at kunne udføre i DK-CLARIN-projektets sidste fase.

3.3. Tilføjelse af manglende information

Hvor tilpasning af hyponymier og sammenlægning af synonyme begreber i teorien ligger nogenlunde lige for (selvom det i praksis er en meget stor opgave at gennemføre det konsistent for hele ordforrådet), så er spørgsmålet om tilføjelse af manglende information i wordnets noget vanskeligere at håndtere. Som det også redegøres for i Pedersen et al. 2009, så er der meget semantisk information i ordbøger for mennesker som tages for givet og derfor ikke ekspliciteres.

For det første ekspliciteres det ikke i definitionerne hvilken type hyponymirelation der er tale om, når der angives et overbegreb. Når det i en DDO-definition angives at *klatretræer* og *egetræer* er *træer*, så er det vigtigt i DanNet at skelne mellem *trætyper* hvor-

til *eg* hører, og træer der snarere opfylder forskellige funktioner (*vejtræer, klatretræer, suttetræer*). Figur 3 viser hvorledes denne skelnen angives i DanNet i og med at kun typer af træer henføres som taksonomiske, hvorimod de andre anføres som ortogonale



(se også Pedersen & Sørensen 2006 for en nærmere redegørelse for denne skelnen).

Figur 3: Skelnen i DanNet mellem taksonomiske begreber og øvrige begreber

En anden oplysning som kun sjældent angives i DDO, er hvem der er den typiske bruger eller frembringer af en genstand. At en maler har malet et billede, at det typisk er en apoteker der driver et apotek, og en læge der bruger en skalpel, og at det typisk er kvinder der bruger stiletter, er ofte oplysninger som regnes for kendte af ordbogsbrugeren eller som må udledes af brugseksempler. Ligeledes angives det heller ikke altid hvad en genstand bruges til; af og til beskrives kun genstandens udseende eller dele, og funk-

tionen skal igen udledes af brugseksemplerne. I DanNet er målet at disse oplysninger angives konsekvent for alle artefakter selvom dette ikke er fuldt gennemført endnu. Nedarvningsmekanismen er en af de funktioner der hjælper DanNet-redaktøren med at identificere sådanne manglende relationer. Hvis der for et professionsbegreb nedarves at en person arbejder, vil man som redaktør blive promptet til at beskrive alle professioner mere specifikt: At en lærer arbejder ved at *undervise*, en læge ved at *behandle* og så fremdeles. En vanskelig opgave i denne henseende er grænsedragningen mellem lingvistisk viden og mere generel omverdensviden som jo i princippet er uendelig og ikke kan rummes i et wordnet. Hvor sætter man grænsen for hvad der skal med? Pustejovskys qualiastruktur med de semantiske relationer den bringer med sig (som fx hvordan noget er blevet til, og hvad det bruges til), taler for at holde sig til den lingvistisk relevante information, og det er i første omgang denne ramme vi forsøger at holde os inden for, svarende til det skel som leksikografer forsøger at drage i forhold til encyklopædisk viden.

4. Anvendelser af semantiske sprogressourcer

Det har tilsyneladende ikke ligget helt lige for at anvende DanNet i kommercielle produkter; man kan sige at markedet har ladet vente på sig. På nuværende tidspunkt har vi kendskab til at DanNet anvendes i OpenOffices danske skrivehjælp som en udvidet synonymifunktion hvor det udnyttes at man kan fremfinde ikke kun "rigtige" synonymer men også nært beslægtede begreber. Samt som en ekstra semantisk visualiseringsfunktion i DDO's egen opslagsfacilitet på nettet (ordnet.dk) hvor man gøres bekendt med beslægtede ord til et givent opslagsord.

DanNet har i skrivende stund været open source i et års tid (siden marts 2009), og vi håber at det blot er et spørgsmål om tid før

flere virksomheder får øjnene op for anvendelsespotentialet. Ved IT- og Telestyrelsens konkurrence i januar 2010 om offentlige data i spil søgte to danske virksomheder om støtte til at eksperimentere med DanNet til forskellige anvendelser, bl.a. til fleksibel søgning i offentlige databaser. Desværre vandt ingen af dem konkurrencen. Vores formodning er at det er en tidskrævende proces for virksomhederne at undersøge potentialet af ressourcen.

På Københavns Universitet har vi de sidste tre år afholdt kurser i sprogteknologi og informationssøgning hvor vi anvender DanNet til semantisk udvidelse af søgeforespørgsler. Ressourcen er afprøvet på tre forskellige domæner (ernæring, uddannelse, biler), og på alle tre tekstsamlinger viser de studerendes eksperimenter en forbedring af "recall" (systemets evne til at udtrække relevante dokumenter) – uden en tilsvarende kritisk forværring af "precision" (systemets evne til at afvise irrelevante dokumenter).

Til forskningsformål anvendes ressourcen på flere niveauer. Der er et lingvistisk anvendelsespotentiale for ressourcen idet databaseformatet gør det muligt at udtrække statistiske oplysninger om leksikalsk semantik på dansk. Det er således muligt at drage mere eller mindre dristige konklusioner på basis heraf: fx at kvinder primært vurderes på deres udseende og seksuelle tilbøjeligheder, mens mænd primært vurderes på deres opførsel. Denne kontroversielle slutning drages i Braasch & Pedersen (2010) ved at undersøge antallet af værdiladede benævnelser for hhv. mænd og kvinder (fx *tøjte* og *vatpik*) og udlede statistik for hvilke egenskaber disse begreber hver især specificerer (seksualitet, opførsel, udseende, formåen mv.). Eller man kan udlede noget om dansk madkultur gennem tiden ved at udtrække to konkurrerende taksonomier for *oste*: Én baseret på en traditionel opfattelse af *ost* som pålæg i form af enten *skæreost* eller *smøreost*, og en anden (eller flere!) baseret på importerede italienske, franske og spanske oste som anvendes i madlavningen på en anden måde. Eller man kan udtrække en stor mængde udtryk for talehandlinger på dansk.

Mulighederne er mange.

Hvordan ressourcen på længere sigt vil indgå i sprogteknologiske forskningsprojekter vil fremtiden vise. Jeg har nævnt informationssøgning, og under dette felt bør også nævnes semantisk annotering af tekst til brug for søgning, resumering og data mining. Semantisk annotering er et af de felter som vi på Center for Sprogteknologi og Det Danske Sprog- og Litteraturselskab gerne vil udforske i fremtiden hvis vi kan opnå fornøden bevilling til opgaven. I den forbindelse skal det afprøves i hvor høj grad DanNets detaljerings- og dækningsgrad er passende til opmærkning af løbende ord i dansk tekst, og i hvor høj grad den rigtige betydning af et ord i en given kontekst kan udledes automatisk ud fra håndopmærkede træningskorpora.

5. Konkluderende bemærkninger

Ressourcer som DanNet er med til at påvise at der er en tæt synergi imellem sprogteknologi og leksikografi idet fællesmængden af relevant viden er stor. Det faktum at moderne leksikografi via nyere korpus- og redigeringsværktøjer har udviklet sig drastisk de seneste tiår og nu producerer mere konsistente og korpusnære resultater, har nok generelt styrket tendensen til at de to felter er rykket nærmere hinanden. Jeg tror yderligere at sprogteknologiske værktøjer som DeepDict (Bick 2009 og i dette bind) og Sketch Engine (Kilgarriff et al. 2004) der ud fra store analyserede korpora genererer typiske leksikalske mønstre, vil styrke denne tendens i de kommende år således at det i fremtiden vil virke besynderligt at udarbejde sprogteknologiske ressourcer uafhængigt af allerede eksisterende leksikografisk materiale og vice versa. Og der er ingen tvivl om at begge felter vil have gavn af denne udvikling.

Litteratur

Bick, E. 2009: *DeepDict – A Graphical Corpus-based Dictionary of Word Relations*. I: *Proceedings of NODALIDA 2009. NEALT Proceedings Series Vol. 4*. Tartu: Tartu University Library, 268-271.

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

Braasch, A. & B. S. Pedersen 2010: Encoding Attitude and Connotation in Wordnets. *Proceedings of the 14th EURALEX Conference*, Leeuwarden, Holland.

Calzolari, N. 1988: The dictionary and the thesaurus can be combined. I: M. Evens: *Relational models of the lexicon. Studies in natural language processing*, Cambridge: Cambridge University Press, 75-96.

Cann, R. 1993: Formal semantics: an introduction. Cambridge University Press.

Copestake, A. 1990: An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. I: *Proceedings of the First International Workshop on Inheritance in Natural Language Processing*, Tilburg, 19-29.

Copestake, A. & T. Briscoe 1995: Semi-productive Polysemy and Sense Extension. I: *Journal of Semantics, Vol 12, 1,*15-67.

Cruse, D.A. 2002: Hyponymy and Its Varieties. I: R. Green, Bean, C.A. & Myaeng, S. H. (eds.), *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*. Springer Verlag, 2-21.

DDO = Hjorth, E., Kristensen, K. et al. (eds.). (2003-2005). *Den Danske Ordbog 1-6*. Gyldendal & Det Danske Sprog- og Litteraturselskab. www.ordnet.dk/ddo.

Fellbaum, C. (ed.) 1998: *WordNet – An Electronic Lexical Database*. Cambridge, Massachusetts, London, England: The MIT Press.

Ide, N. & J. Véronis 1995: Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation. I: Steffens, P. (ed.), *Machine Translation and the Lexicon, Third International EAMT Workshop*, *Heidelberg 1993, Proceedings*. Lecture Notes in Computer Science 898. Springer.

Ide, N. & Y. Wilks 2007: Making Sense About Sense. I: E. Agirre & P. Edmonds eds.) *Word Sense Disambiguation – Algorithms and Applications*. Springer, 47-75.

Kaplan, R. M. & J. Bresnan 1982: Lexical Functional Grammar: A Formal System for Grammatical Representation. I: J. Bresnan: *Mental Representation of Grammatical Relations*, MIT Press.

Kilgarriff, A. 1997: I don't believe in word senses. I: *Computers and the Humanities 31,* 91-113.

Kilgarriff, A., P. Rychly, P. Smrz 2004: The Sketch Engine. *Proceedings of EURALEX 2004*, Lorient, France, 105-116.

Kokkinakis, D., M. Toporowska Gronostaj M. & K. Warmenius 2000: Annotating, Disambiguating & Automatically Extending the Coverage of the Swedish SIMPLE Lexicon, Språkdata, Göteborg. I: *Proceedings of the 2nd LREC (Language Resources and Evaluation Conference)*, Athens, Hellas, 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 2000 13(4)*, 249-263.

Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen 2009: DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation vol. 43*, 269-299.

Pedersen, B.S, S. Nimb, A. Braasch 2010: Merging specialist taxonomies and folk taxonomies in wordnets – a case study of plants, animals and foods in the Danish wordnet. I: *Proceedings from Language Resources and Evaluation Conference*. Malta.

Pedersen, B., Paggio, P. 2004: The Danish SIMPLE Lexicon and its Application in Content-based Querying. I: *Nordic Journal of Linguistics Vol 27:1*, 97-127.

Pedersen, B.S., N. Sørensen 2006: Towards Sounder Taxonomies in Wordnets. I: A. Oltramari, Chu-Ren Huang, A. Lenci, P. Buitelaar, C. Fellbaum (eds): *Ontolex 2006 at 5th International Conference on Language Resources and Evaluation*. Genova, Italy, 9-16.

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

Prebensen, H. 2006: Formens fascination. I: A. Braasch, C. Navarretta, S. Nimb, S. Olsen, P. Paggio, B. Pedersen (red.): *Sprogteknologi i dansk perspektiv – En samling artikler om sprogforskning og automatisk sprogbehandling*. København: Reitzels Forlag, 51-69.

Pustejovsky, J. 1995: *The Generative Lexicon*. MIT Press.

Spang-Hanssen, E. 2006: Sprogteknologi og humaniora. I: A. Braasch, C. Navarretta, S. Nimb, S. Olsen, P. Paggio, B. Pedersen (red.): *Sprogteknologi i dansk perspektiv – En samling artikler om sprogforskning og automatisk sprogbehandling*. København: Reitzels Forlag, 39-51.

Wierzbicka, A. 1996: *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Bolette Sandford Pedersen
professor, ph.d.
Københavns Universitet
Njalsgade 140
DK-2300 København S
bspedersen@hum.ku.dk

Sprogteknologiske ressourcer for islandsk leksikografi

Eiríkur Rögnvaldsson

Ten years ago, the Icelandic government launched a special Language Technology Program with the aim of supporting institutions and companies in creating basic resources for Icelandic language technology work. This initiative resulted in the creation and development of several important resources and tools that have had profound influence on Icelandic language technology, and are also valuable for Icelandic lexicography and linguistic research in general. The present paper describes briefly some of the most important of these products, such as a morphological database (260,000 lemmas), a 25 million word balanced PoS tagged corpus, a lemmatiser, a rule-based tagger, and a shallow parser. Finally, it is pointed out that all the tools that the Icelandic Language Technology Community has developed in the past few years have been made Open Source, and the importance of adopting Open Source Policy for small language communities is emphasized.

1. Indledning

Denne artikel handler om de vigtigste sprogteknologiske ressourcer som eksisterer for islandsk.[1] Selv om disse ressourcer oprindelig blev lavet for at benyttes i sprogteknologiske værktøjer, så kan de fleste af dem også være nyttige for leksikografien. Der begyndes med at opridse baggrunden for at de fleste af disse ressourcer blev lavet – det islandske sprogteknologiprojekt som begyndte i 2001. Derefter beskrives de vigtigste af disse ressourcer kort – en morfo-

1 Mange tak til redaktørerne for *LexicoNordica* for kommentarer og rettelser.

logisk database, et balanceret korpus og sprogteknologiske værktøjer som en lemmatiser, en morfologisk tagger og en syntaktisk parser. Til slut berøres spørgsmålet om open source policy som forfatteren anser for at være meget vigtig for et lille sprogsamfund som det islandske.

2. Islandsk sprogteknologi omkring århundredskiftet

Man kan godt sige at islandsk sprogteknologi simpelthen ikke eksisterede for ti år siden. Der fandtes ganske vist et godt stavekontrolsystem og en talesyntese, omend den var relativt primitiv. Men det var alt. Der eksisterede ingen programmer eller enkelte kurser i sprogteknologi eller datalingvistik ved islandske universiteter eller højskoler, og der fandtes ikke nogen som bedrev akademisk forskning inden for dette felt. Der var heller ikke nogen private firmaer som arbejdede med sprogteknologi.

Dette har nu ændret sig. I efteråret 1998 oprettede ministeren for undervisning og forskning et udvalg som skulle undersøge situationen for islandsk sprogteknologi og komme med forslag til hvordan man kunne styrke islandsk sprogteknologi. Der var tre medlemmer i udvalget: Rögnvaldur Ólafsson, docent i fysik, Eiríkur Rögnvaldsson, professor i islandsk sprog, og Þorgeir Sigurðsson, ingeniør og lingvist.

Udvalget leverede sin rapport til ministeren i april 1999 (Ólafsson m.fl. 1999), og to år senere, i år 2001, oprettede regeringen et specielt sprogteknologiprojekt (Arnalds 2004, Ólafsson 2004). Projektets hovedmål var at give institutioner og firmaer støtte til at opbygge grundressourcer for islandsk sprogteknologi. Dette initiativ har ført til oprettelsen af forskellige projekter som har haft stor indflydelse på dette område.

Sprogteknologiudvalget foreslog fire typer af aktiviteter til at styrke islandsk sprogteknologi (Ólafsson m.fl. 1999):

- Sproglige ressourcer skulle udvikles og opbygges til anvendelse for firmaer som ville udvikle sprogteknologiske værktøjer og andre produkter.
- Praktisk forskning inden for sprogteknologi skulle støttes.
- Firmaer skulle støttes til at udvikle sprogteknologiske produkter.
- Universitetsprogrammer og kurser i sprogteknologi skulle oprettes.

Alle disse aktiviteter er blevet gennemført, i det mindste til en vis grad (Arnalds 2004, Ólafsson 2004, Rögnvaldsson 2005). Vigtige ressourcer er blevet opbygget, sprogteknologisk forskning er blevet igangsat, støtte til at udvikle sprogteknologiske værktøjer er blevet givet til firmaer, og en tværfaglig mastergrad i sprogteknologi er blevet oprettet i samarbejde mellem Islands Universitet og Reykjavik University. De vigtigste ressourcer som blev udviklet inden for sprogteknologiprojektet, er følgende:

- En morfologisk database med omkring 260.000 ord.
- Et balanceret og tagget korpus med 25 millioner ord.
- En statistisk PoS-tagger.
- Islandsk talesyntese.
- Islandsk talegenkender.
- Et forbedret stavekontrolsystem.

I det følgende gives der en kort beskrivelse af de af disse ressourcer som kan være til nytte for islandsk leksikografi (se også Rögnvaldsson 2008, Rögnvaldsson m.fl. 2009).

3. Sprogteknologiske ressourcer

3.1. Morfologisk database

En af de vigtigste ressourcer som blev sat i gang under regeringens sprogteknologiprojekt, er en morfologisk database for islandsk sprog, Beygingarlýsing íslensks nútímamáls (bøjningsbeskrivelse af islandsk nutidssprog, BÍN; Bjarnadóttir 2005) som der arbejdes på ved Leksikografisk institut ved Islands Universitet, der nu er blevet en del af Árni Magnússon-instituttet for islandske studier. Projektet blev påbegyndt i 2002 og var i begyndelsen finansieret af den islandske sprogteknologifond. Den første fase af projektet blev afsluttet i 2004, men arbejdet fortsættes – nu finansieret af instituttet. Kristín Bjarnadóttir har været projektleder fra begyndelsen. Databasen indeholder nu paradigmer for næsten 260.000 ord, mens der er mere end 5,6 millioner ordformer.

Databasen blev oprindelig oprettet med to formål. Først og fremmest skulle den kunne bruges i forskellige sprogteknologiske produkter og værktøjer. Det var faktisk hovedgrunden til at den blev lavet, for uden finansiering fra sprogteknologifonden havde det været umuligt at sætte projektet i gang. Men databasen har også været til stor nytte for almene brugere efter at den er kommet på internettet og alle kan slå bøjningen af islandske ord op i den. Denne mulighed er faktisk blevet meget populær, og både skoleelever og andre bruger databasen meget.

Databasen er allerede blevet benyttet i nogle praktiske produkter, særlig søgeprogrammer. Når man bruger disse programmer til søgning efter islandske ord, behøver man ikke indtaste alle mulige bøjningsformer fordi programmerne har adgang til databasen og derfor automatisk kan søge efter alle former – også de uregelmæssige. Databasen bruges fx i telefonbogen på internettet (http://ja.is) hvilket betyder at man kan taste hvilken som helst bøjnings-

form af et givent navn ind – telefonbogen vil vise navnene i den almindelige opslagsform (dvs. nominativ). Databasen bruges også i læremateriale*t Icelandic Online* på internettet (http://icelandic.hi.is). Desuden bruges den i forskellige projekter der vedrører tagging og lemmatisering af islandske tekster.

Databasen var oprindelig tekstbaseret og lagret som XML-filer, men nu er den lige blevet omstruktureret og filerne lagt ind i en MySQL-database. Kristín Bjarnadóttir har været ansvarlig for den lingvistiske del af omstruktureringen, mens Hjálmar Gíslason har været ansvarlig for al programmering. Formålet med omstruktureringen er på den ene side at gøre det nemmere at føje nye ord til databasen og rette fejl i den og på den anden side at gøre søgning i databasen hurtigere og mere effektiv.

3.2. Balanceret tagget korpus

En anden ressource som er blevet udviklet med støtte fra sprogteknologiprojektet, er et balanceret PoS-tagget korpus på omkring 25 millioner ord (Helgadóttir 2005, 2009; Loftsson m.fl. 2010). Dette korpus er også blevet udviklet på Leksikografisk institut, og projektlederen er statistikeren Sigrún Helgadóttir. Projektet blev påbegyndt i 2004 og skal være færdigt i 2011. Indsamlingsfasen var meget sværere og langsommere end man havde forventet. Grunden hertil var hovedsagelig problemer angående ophavsret. Det tog vældig lang tid at få lov hos alle ophavsretsindehavere til at bruge deres tekster. Det viste sig også at bogforlagene var vældig skeptiske over for projektet og ikke særlig villige til at udlevere de tekster som de var i besiddelse af. Men til sidst lykkedes det at forhandle en aftale med forlagene på plads, og indsamlingen af tekster er nu færdig.

Målet var at få fat i så mange teksttyper som muligt, og der findes i alt omkring 20 forskellige teksttyper i korpusset. Størstedelen af materialet kommer fra de følgende teksttyper: Avistekster

(28%), trykte bøger (romaner osv.) (22%), blog (7,5%), forskellige tidsskrifter (8,5%), tekst fra den islandske "Videnskabsweb" (http://visindavefur.hi.is) (7,5%), webtekster fra institutter, firmaer, foreninger etc. (6%), love og andre tekster fra Altinget (3%) samt talesprog (2%) (Helgadóttir 2009; se også Loftsson m.fl. 2010). Størstedelen af talesprogsmaterialet er samtaler fra projektet Ístal – islandsk talesprogsbank – og blev indsamlet i år 2000. Alle teksterne er fra det 21. århundrede, dvs. årene 2000–2009.

I løbet af 2010 bliver teksterne PoS-tagget og lagret i TEI-kompatibelt XML-format (http://www.tei-c.org/release/doc/tei-p4-doc/html/). Nu arbejdes der på en brugergrænseflade for korpusset. Det skal være søgbart gennem internettet, og man skal kunne søge efter enkelte ord og ordforbindelser, og også efter tags – fx alle substantiver i femininum, singularis, dativ. Projektet har fået adgang til søgesystemet Glossa hos Tekstlaboratoriet i Oslo og er i gang med at tilpasse det til korpusset. Dette arbejde vil forhåbentlig blive færdigt i begyndelsen af 2011, men en præliminær version findes allerede på webben (http://mim.hi.is). Så vil alle kunne søge i korpusset, men af ophavsretlige grunde vil man kun se et brudstykke af hver tekst ad gangen. De som har brug for hele korpusset, enten til lingvistisk forskning eller for at udvikle sprogteknologiske værktøjer, vil kunne få hele korpusset udleveret, men må i givet fald underskrive en kontrakt. Dette er en model der kendes andre steder fra, og ligner fx den som bruges til den danske orddatabase STO.

Dette korpus vil forhåbentlig blive af stor nytte for alle som arbejder med islandsk sprog og islandsk leksikografi. Det vil blive muligt at få meget bedre information end man har haft hidtil om ordfrekvens, forskellen mellem teksttyper, nye ord, ændringer i bøjningen osv. Korpusset er halvtreds gange så stort som det største taggede korpus man har haft hidtil, nemlig det som *Íslensk orðtíðnibók* (islandsk frekvensordbog, Pind m.fl. 1991) er baseret på, og det indeholder mange teksttyper som ikke fandtes i basen til frekvensordbogen. Særlig værdifuldt er det at have tekster af

uformelt register, såsom blog og talesprog. I disse tekster kan man finde mange sprogændringer som ikke forekommer i de mere formelle tekster og ikke hidtil har været registreret i ordbøger. Dette handler naturligvis om nye ord, både låneord og nydannelser, men også om ændringer i ordenes betydning og brug, ændringer i bøjningen osv.

Til tagningen bliver der brugt et detaljeret tagsæt som giver brugerne mulighed for at bruge teksten til syntaktisk forskning som også kan benyttes ved leksikografisk beskrivelse. Tagsættet er ganske vist morfologisk baseret og ikke syntaktisk, men da der er en nær forbindelse mellem morfologien og syntaksen, kan man også hente en del syntaktiske oplysninger i de morfologiske tags (se fx Rögnvaldsson og Helgadóttir 2008).

3.3. Andre sproglige ressourcer

To andre ressourcer som er blevet lavet inden for andre projekter som støttedes af regeringens sprogteknologiprogram, bør også nævnes. Den ene er en ordliste med lydskrift. I 2003 gik Islands Universitet og nogle private firmaer sammen om at lave en islandsk talegenkendelse. En vigtig del af projektet var at lave en fonetisk ordliste for islandsk. Der blev lavet en ordliste som indeholder 56.000 af de hyppigste ordformer i islandsk. Listen byggede på frekvenslister som blev sammenstillet ud fra forskellige tekster som blev stillet til rådighed for projektet – avistekster fra Islands største avis, *Morgunblaðið*, omkring 100 bøger fra Islands største forlag og flere mindre tekstsamlinger. Fire studerende inden for lingvistik og sprogteknologi fik til opgave at transskribere listen med lydsskriftsystemet SAMPA. Senere er transskriptionen blevet konverteret til IPA, så nu findes der en liste med to typer af fonetisk transskription. Denne liste er allerede blevet benyttet både til islandsk talegenkendelse og islandsk talesyntese. Den kan naturligvis også bruges til traditionelle ordbøger.

Den anden ressource er resultatet af et syntaktisk parsing-projekt som desværre ikke blev afsluttet. Det er en liste på 28.000 linjer med verber og deres argumentstruktur. Listen viser hvilke argumenter hvert verbum kan have, i hvilke kasus argumenterne står, hvilke præpositioner verberne tager osv. Hvert verbum får ofte mange linjer da mange verber kan have flere forskellige argumentstrukturer. Listen er lavet ud fra eksisterende ordbøger, og den kan naturligvis være af stor nytte i leksikografisk arbejde.

4. Islandsk sprogteknologi i dag

4.1. Sprogteknologiske værktøjer

Foruden de sproglige ressourcer som der er blevet gjort rede for ovenfor, er der blevet udviklet nogle sprogteknologiske værktøjer som naturligvis også kan være af stor nytte for leksikografien. Heldigvis var originalfilerne fra Islandsk frekvensordbog blevet opbevaret på Leksikografisk institut og kunne bruges til at træne statistiske taggere. Den første tagger som blev trænet ved brug af listerne, var μ-tbl (Lager 1999), men den tagger som gav det bedste resultat, var TnT (Brants 2000). Den bedste score man har fået, er 92% korrekt (Helgadóttir 2007). Det er vistnok ikke særlig godt hvis man sammenligner det med resultater for mange andre sprog, som engelsk eller svensk, men tagsættet er også usædvanlig stort – omkring 700 forskellige tags.

Hrafn Loftsson, lektor ved Reykjavik Universitet, har udviklet en regelbaseret tagger, IceTagger, som giver lidt bedre resultat end TnT (Loftsson 2008). Hrafn og flere har også eksperimenteret med forskellige kombinationer af taggere og forenkling af tagsættet (Henrich m.fl. 2009). Dette giver op til 93,70% korrekt tagging.

Hrafn har også skrevet en syntaktisk parser, IceParser (Loftsson og Rögnvaldsson 2007). Dette er en såkaldt "shallow parser"

som ikke udfører en fuld syntaktisk analyse, men genkender de vigtigste syntaktiske fraser, såsom nominalfraser, præpositionsfraser osv. Parseren markerer også de vigtigste syntaktiske funktioner, såsom subjekt, objekt osv. IceParser udfører en analyse som minder om Constraint Grammar (CG, se Karlsson 1990), men den er dog ikke CG-baseret. Hrafn har imidlertid planer om at omskrive sin parser så at den bliver ren CG-parser.

Indtil for nylig har man ikke haft nogen lemmatiser for islandsk. Der er blevet eksperimenteret med at træne CST's lemmatiser (http://cst.dk/online/lemmatiser/) på islandsk tekst, og det gav et godt resultat. Anton Karl Ingason, som er studerende inden for sprogteknologi og lingvistik, har nu skrevet en lemmatiser for islandsk, Lemmald, som giver lidt bedre resultater (Ingason m.fl. 2008).

Alle disse værktøjer er nu online på http://nlp.cs.ru.is. De kan bruges én ad gangen eller alle samtidig (dvs. man kan få teksten PoS-tagget, parset og lemmatiseret på én gang).

4.2. Nuværende projekter

Efter at regeringens sprogteknologiprojekt blev afsluttet i slutningen af 2004, så det ud til at udviklingen af sprogteknologiske værktøjer for islandsk ikke ville kunne fortsættes. Men heldigvis havde man på dette tidspunkt fået samlet en solid gruppe forskere og studerende fra tre institutioner – Islands Universitet, Reykjavik University og Árni Magnússon-instituttet for islandske studier. Denne gruppe har stået bag næsten alle sprogteknologiske projekter på Island i løbet af de sidste ti år.

I begyndelsen af 2009 fik sprogteknologigruppen tildelt et stort treårs stipendium fra den islandske forskningsfond til projektet *Viable Language Technology beyond English – Icelandic as a Test Case* (http://iceblark.wordpress.com). Formålet med dette projekt er at fortsætte med at opbygge ressourcer for islandsk sprogtekno-

logi på den billigste og mest effektive måde og derved bidrage til en islandsk såkaldt BLARK (Basic Language Resource Kit, se Krauwer 2003). Projektet har tre delprojekter. Ét er det semantiske projekt som Anna Björk Nikulásdóttir og Matthew Whelpton (2010) beskriver i deres artikel i dette nummer af *LexicoNordica*. Et andet er maskinoversættelse fra islandsk til engelsk ved hjælp af Apertium-programmet der er udviklet ved universitetet i Alicante (http://www.apertium.org/). En prøveudgave af oversættelsessystemet er nu tilgængelig på http://nlp.cs.ru.is/ApertiumISENWeb/.

Det tredje og største delprojekt er en træbank, dvs. et syntaktisk analyseret korpus. Ved opbygningen af korpusset samarbejder projektet med en gruppe fra University of Pennsylvania som har stået bag nogle af de største og bedst kendte projekter af denne slags – Penn Treebank (Marcus m.fl. 1993) og Penn Parsed Corpora of Historical English (PPCME2, PPCEME). Det er kostbart og tidkrævende at opbygge en træbank, og derfor er det meget vigtigt at automatisere processen så vidt som muligt. Der eksperimenteres p.t. med at bruge IceTagger og IceParser til præliminær analyse og koble dem til programmer fra sprogteknologigruppens medarbejdere i University of Pennsylvania. En præliminær version af træbanken (IcePaHC) er allerede blevet lagt ud på internettet og kan downloades fra http://linguist.is/icelandic_treebank.

5. Konklusion

I denne artikel er der blevet gjort rede for de vigtigste sprogteknologiske ressourcer og værktøjer som findes for islandsk. Men det er ikke nok at disse ting eksisterer; et stort og velkendt problem ved mange sproglige ressourcer at de er kommercielle og må købes – ofte til en høj pris. Da den islandske regerings sprogteknologiprojekt begyndte i 2001, blev det samtidig besluttet at staten skulle finansiere opbygningen af forskellige ressourcer for islandsk

sprogteknologi da det islandske marked er alt for lille til at kommercielle firmaer har råd til at bygge sådanne ressourcer op.

Det blev desværre ikke besluttet fra begyndelsen at disse ressourcer skulle være gratis, kun at alle som ville bruge dem til sprogteknologiske projekter skulle kunne få dem til rimelig pris. Man håbede at de penge som firmaer ville betale for at få lov til at benytte dem, ville være nok til at vedligeholde dem. Det har dog vist sig at det er vældig svært at finde ud hvad der er en rimelig pris, og at selv en lav pris er en stor tærskel for benyttelse af ressourcerne. Der er også mange som gerne ville eksperimentere med ressourcerne, men ikke kan betale for dem og ikke bryder sig om eller finder det værd at læse og underskrive alle mulige kontrakter for at kunne bruge dem til videnskabelige formål. Dette har ført til at ressourcerne ikke er brugt så meget som man havde håbet.

Derfor har den islandske sprogteknologigruppe indset at det er nødvendigt at alle sproglige og sprogteknologiske ressourcer for islandsk bliver open source i videst muligt omfang. Dette er allerede blevet gjort ved de fleste af de sprogteknologiske værktøjer som medlemmer af gruppen har lavet. Taggeren IceTagger, parseren IceParser og lemmatiseringsprogrammet Lemmald indgår alle i programpakken IceNLP som er licenseret under GNU LGPL (Lesser General Public License, se http://www.gnu.org/licenses/lgpl.html) og findes på http://sourceforge.net/projects/icenlp.

Det er klart at disse ressourcer og værktøjer kan blive af stor nytte for leksikografien. Men der mangler endnu forskellige grundlæggende ressourcer, ikke mindst en god sprogteknologisk orddatabase som fx den danske STO, med morfologiske, syntaktiske og semantiske oplysninger. Der eksisterer faktisk meget materiale til en orddatabase af denne slags, og man kan vel sige at en af de vigtigste opgaver for islandsk sprogteknologi i de næste år, er at finde en måde at koble alle disse ressourcer sammen. Det ville åbne nye og fascinerende muligheder for islandsk leksikografi.

Litteratur

Arnalds, Ari 2004: Language Technology in Iceland. I: Henrik Holmboe (red.): *Nordisk Sprogteknologi. Årbog 2003*. København: Museum Tusculanums Forlag, 41–43.

Bjarnadóttir, Kristín 2005: Modern Icelandic Inflections. I: Henrik Holmboe (red.): *Nordisk Sprogteknologi. Årbog 2005*. København: Museum Tusculanums Forlag, 49–50.

Brants, Thorsten 2000: TnT – A Statistical Part-of-Speech Tagger. I: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*. Seattle, 224–231.

Henrich, Verena, Timo Reuter og Hrafn Loftsson 2009: CombiTagger: A System for Developing Combined Taggers. I: *Proceedings of the 22nd International FLAIRS Conference, Special Track: "Applied Natural Language Processing"*. Sanibel Island, Florida. http://aaai.org/ocs/index.php/FLAIRS/2009/paper/view/67/296

Helgadóttir, Sigrún 2004: Mörkuð íslensk málheild. I: *Samspil tungu og tækni*. Reykjavík: Ministeriet for undervisning, forskning og kultur, 67–71.

Helgadóttir, Sigrún 2007: Mörkun íslensks texta. I: *Orð og tunga 9*, 75–107.

Helgadóttir, Sigrún 2009: Mörkun texta og markaðar málheildir. Foredrag på Hugvísindaþing, Reykjavík, 14. marts.

Ingason, Anton Karl, Sigrún Helgadóttir, Hrafn Loftsson og Eiríkur Rögnvaldsson 2008: A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). I: Bengt Nordström og Aarne Ranta (red.): *Advances in Natural Language Processing*. Lecture Notes in Computer Science, Vol. 5221. Berlin: Springer, 205–216.

Karlsson, Fred 1990. Constraint Grammar as a Framework for Parsing Unrestricted Text. I: Hans Karlgren (red.): *Proceedings*

of the 13th International Conference of Computational Linguistics, Vol. 3. Helsingfors, 168–173.

Krauwer, Steven 2003: The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. I: *Proceedings of SPECOM 2003.* Moskva, 8–15.

Lager, Torbjörn 1999: The μ-TBL System: Logic Programming Tools for Transformation-Based Learning. I: *Proceedings of the Third International Workshop on Computational Natural Language Learning* (CoNLL'99), Bergen. http://www.cnts.ua.ac.be/conll99/programme.html

Loftsson, Hrafn 2008: Tagging Icelandic text: A linguistic rule-based approach. I: *Nordic Journal of Linguistics 31*, 47–72.

Loftsson, Hrafn, og Eiríkur Rögnvaldsson 2007: IceParser: An Incremental Finite-State Parser for Icelandic. I: Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek og Mare Koit (red.): *Proceedings of the 16th Nordic Conference of Computational Linguistics*. Tartu, 128–135.

Loftsson, Hrafn, Jökull H. Yngvason, Sigrún Helgadóttir og Eiríkur Rögnvaldsson 2010: Developing a PoS-tagged corpus using existing tools. I: *Proceedings of "Creation and use of basic lexical resources for less-resourced languages", workshop at the 7th International Conference on Language Resources and Evaluation, LREC 2010*. Valetta, 53–60.

Marcus, Mitch, Beatrice Santorini og Mary Ann Marcinkiewicz 1993: Building a Large Annotated Corpus of English: The Penn Treebank. I: *Computational Linguistics 19(2)*, 313–330.

Nikulásdóttir, Anna Björk og Matthew Whelpton 2010: Lexical Acquisition through Noun Clustering. I: *LexicoNordica 17* (i dette bind).

Ólafsson, Rögnvaldur 2004: Tungutækniverkefni menntamálaráðuneytisins. I: *Samspil tungu og tækni*. Reykjavík: Ministeriet for undervisning, forskning og kultur, 7–13.

Ólafsson, Rögnvaldur, Eiríkur Rögnvaldsson og Þorgeir Sigurðsson 1999: *Tungutækni. Skýrsla starfshóps.* Reykjavík: Ministeriet for undervisning, forskning og kultur.

Pind, Jörgen (red.), Friðrik Magnússon og Stefán Briem 1991: *Íslensk orðtíðnibók.* Reykjavík: Orðabók Háskólans.

Rögnvaldsson, Eiríkur 2005: Staða íslenskrar tungutækni við lok tungutækniátaks *Tölvumál*, 24. februar.

Rögnvaldsson, Eiríkur 2008: Icelandic Language Technology Ten Years Later. I: *Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages.* SALTMIL workshop, LREC 2008. Marrakech, 1–5.

Rögnvaldsson, Eiríkur og Sigrún Helgadóttir 2008: Morphological Tagging of Old Norse Texts and Its Use in Studying Syntactic Variation and Change. I: *2nd Workshop on Language Technology for Cultural Heritage Data.* LREC 2008 workshop. Marrakech, 40–46.

Rögnvaldsson, Eiríkur, Hrafn Loftsson, Kristín Bjarnadóttir, Sigrún Helgadóttir, Anna Björk Nikulásdóttir, Matthew Whelpton og Anton Karl Ingason 2009: Icelandic Language Resources and Technology: Status and Prospects. I: Rickard Domeij, Kimmo Koskenniemi, Steven Krauwer, Bente Maegaard, Eiríkur Rögnvaldsson og Koenraad de Smedt (red.): *Proceedings of the NODALIDA 2009 workshop Nordic Perspectives on the CLARIN Infrastructure of Language Resources*. Tartu: Northern European Association for Language Technology (NEALT), Tartu University Library, 27–32.

Internethenvisninger

BÍN = Bjarnadóttir, Kristín (red.): *Beygingarlýsing íslensks nútímamáls.* http://bin.arnastofnun.is

Icelandic Online. Undervisningsmateriale i islandsk for begyndere. http://icelandic.hi.is/

IcePaHC = Wallenberg, Joel, Anton Karl Ingason, Einar Freyr Sigurðsson og Eiríkur Rögnvaldsson 2010. *Icelandic Parsed Historical Corpus (IcePaHC)*. Version 0.1. http://linguist.is/icelandic_treebank

PPCEME = Kroch, Anthony, Beatrice Santorini og Lauren Delfs 2004: *Penn-Helsinki Parsed Corpus of Early Modern English.* http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/

PPCME2 = Kroch, Anthony og Ann Taylor 2000: *Penn-Helsinki Parsed Corpus of Middle English, second edition.* http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/

STO = Braasch, Anna m.fl. (red.): *Sprogteknologisk Ordbase.* København: Center for Sprogteknologi 2001–2004. http://cst.dk/sto

Eiríkur Rögnvaldsson
professor
Islands Universitet
Árnagarði við Suðurgötu
IS-101 Reykjavík
eirikur@hi.is

Svenska ordboksredigeringssystem – med fokus på Cronoma

Christian Sjögreen & Emma Sköldberg

Our paper concerns dictionary writing systems (DWS), i.e. software for writing and producing dictionaries. We discuss some necessary features of a DWS and the use of different DWSs in Sweden and, to some extent, also internationally. Given that DWSs are expensive to develop and maintain, the DWS situation in Swedish publishing houses is rather unsatisfactory. More efficient tailor-made systems have been developed at academic lexicographic institutes in Sweden. One such system is *Cronoma*, a DWS used to compile *Lexin Svenska ord* (Lexin Swedish words) in the Lexin project at the Centre for Lexicology and Lexicography at the University of Gothenburg. Cronoma has been developed in close co-operation between lexicographers and systems engineers. It is considered easy to use and will soon be available as a free resource.

1. Inledning

För lexikografer är det oerhört viktigt att ha tillgång till bra tekniska verktyg för att hantera alla de data som ligger till grund för en ordbok. Ett sådant verktyg är själva ordboksredigeringssystemet. Atkins/ Rundell (2008:114–117) menar att ett typiskt ordboksredigeringssystem har tre huvudkomponenter: a) ett textredigeringssystem, i vilket lexikografen skapar och redigerar ordboksartiklar, b) en databas, i vilken den framväxande ordbokstexten lagras och c) en uppsättning administrativa verktyg, vilka stöder hanteringen av projektet och publiceringsprocessen. De flesta lexikografer är väl förtrogna med det eller de system som används inom den egna redaktionen. Vissa system har också demonstrerats, t.ex. i

samband med internationella workshops om ordboksredigeringssystem i Brighton 2002 och 2003, Brno 2004 och Turin 2006. I övrigt är dock litteraturen begränsad om hur de system som används i Norden faktiskt fungerar.

Vår artikel är ett bidrag till att täcka denna kunskapslucka. För det första kommer vi mer generellt att diskutera redigeringssystem för ordböcker och de krav som ställs på dem. För det andra kommer vi att redogöra för en mindre undersökning av vilka system som i nuläget används i Sverige och i viss mån utanför Sverige. Sista delen av artikeln ägnas åt ett visst redigeringssystem, Cronoma, som bl.a. används vid uppdateringsarbetet med den enspråkiga ordboken *Lexin Svenska ord* som nu sker vid Lexikaliska institutet, Institutionen för svenska språket vid Göteborgs universitet. Vi är väl bekanta med systemet – om än på olika sätt; Christian Sjögreen har utvecklat det och Emma Sköldberg arbetar dagligen med det i sitt arbete som lexikograf.

Ett återkommande tema på Schæffergårds-symposiet 2010 var en önskan att göra olika lexikala och lexikografiska resurser fritt tillgängliga. Det finns i nuläget ytterst få redigeringssystem som är fria att ta del av. När väl arbetet med uppdateringen av den svenska Lexin-databasen är avslutat, har vi emellertid för avsikt att göra en version av Cronoma fritt tillgänglig. Med detta i åtanke kan det vara intressant med en kortare presentation av systemet. En bakomliggande tanke är också att artikeln kan tjäna som dokumentation över detta system, som tidigare knappast har presenterats i skrift. Redigeringssystemet kommer framför allt att diskuteras ur lexikografens perspektiv.

2. Krav på ett optimalt ordboksredigeringssystem

Det ställs höga krav på ett redigeringssystem som ska användas för att producera en ordbok. Ett optimalt ordboksredigeringssystem

förenklar och automatiserar mycket av det arbete som tidigare gjordes manuellt, vilket givetvis leder till att lexikografen arbetar snabbare. Det leder också till högre kvalitet på ordboken, dvs. mer korrekta uppgifter och större intern konsekvens inom verket. Ett riktigt bra system är dessutom lättöverskådligt och enkelt för lexikografer att lära sig. Lexikograferna kan därmed fokusera på själva lexikografin istället för på tekniken (se vidare t.ex. Svensén 2004:498–502, Rasmussen 2006, Atkins/Rundell 2008:115–117).

Vidare kan ett idealt system användas såväl vid framställning av nya som vid revidering av äldre ordböcker. Det fungerar också för framställning av både en- och tvåspråkiga ordböcker, när målet är en pappersordbok eller en elektronisk ordbok och vid utväxling av data mellan bl.a. ordboksredaktion och förlag. En konsekvens av dessa möjligheter är att det lexikografiska projektet inte behöver införskaffa olika system beroende på den ordbok som för tillfället är under produktion (Rasmussen 2006:352).

Sist men inte minst måste ordboksredigeringssystemet fungera under en längre tidsperiod. Det finns flera system som inte stått emot tidens tand, vilket vållat problem för olika lexikografiska projekt. Som exempel kan vi nämna det danska redigeringssystemet GestorLEX, som byggde på operativsystemet OS2. När OS2 avvecklades blev de redaktioner som använde detta redigeringssystem tvungna att byta (se t.ex. Rasmussen 2006, se även avsnitt 3.1 nedan). Systemet måste således vara flexibelt både vad gäller befintliga och möjliga framtida behov.

Med tanke på de krav på systemen som tagits upp ovan är det lätt att förstå att ett bra ordboksredigeringssystem är dyrt att underhålla och införskaffa. Den stora kostnaden är ett viktigt skäl till att det finns så få fritt tillgängliga system. Ett system som trots allt är fritt är bl.a. Lexique Pro, som har utvecklats av amerikanska Summer Institute of Linguistics (SIL). Detta system har använts av lingvister och lexikografer som intresserat sig för mindre dokumenterade språk i t.ex. Afrika, Asien och Stillahavsområdet (se vidare Lexique Pro).

3. Ordboksredigeringssystem i Sverige och internationellt

I det följande presenterar vi kortfattat de system som används idag vid framställningen av enspråkiga ordböcker i Sverige.[1] Vi kommer även i någon mån att blicka utanför Sveriges gränser.

3.1. Ordboksredigeringssystem vid svenska förlag

Det största svenska förlaget i ordbokssammanhang är utan tvivel Norstedts. Inom förlaget används sedan 2004 systemet DicSy, en vidareutveckling av Henning Madsens DOS-baserade system Compulexis. DicSy beskrivs närmare i Svensén (2004:500–502). Efter att ha haft tydliga tekniska problem med detta redigeringssystem har dock Norstedts bestämt sig för att byta system och utvärderar – i skrivande stund – olika kommersiella system, bl.a. det danska iLEX (se vidare iLEX). Det är förvisso dyrt att införskaffa ett system, men på längre sikt är det också kostsamt att hålla fast vid ett system som orsakar problem i det dagliga arbetet. Det är också viktigt att kartlägga de behov som finns inom verksamheten när man ska investera i ett nytt system. Ett belysande exempel är danska Gyldendals omfattande kravspecifikation när förlaget sökte efter ett nytt ordboksredigeringssystem (se vidare Rasmussen 2006).

Ett annat förlag i svenska lexikografiska sammanhang är Bonniers. Den första upplagan av *Bonniers svenska ordbok* kom 1980, och den tionde kommer i år. Denna enspråkiga ordbok torde vara den svenska ordbok som sålts i flest exemplar; antalet borde ligga i

1 Vi tackar Mathias Thiel på Norstedts, Peter A. Sjögren på Bonniers, Christian Mattsson, tidigare på Natur och Kultur men numera på Språkrådet, Eva Thelin på Dialektavdelningen vid Institutet för språk och folkminnen i Uppsala samt Annika Kjellandsson vid Institutionen för svenska språket, Göteborgs universitet för deras värdefulla upplysningar.

trakten av 1 miljon. Trots dessa höga försäljningssiffror har förlaget inte satsat på ett specifikt ordboksredigeringssystem utan hela ordboksmaterialet finns i samma sättnings- och sidombrytningssystem som används för vilken annan bok som helst i Bonniers utgivning. Nya artiklar skrivs i Microsoft Word, och befintligt material redigeras på papperskopior.

Vårt tredje och sista exempel är bokförlaget Natur och Kultur, där det redan nämnda GestorLEX användes mellan ca 1991 och 2006. Eftersom lexikonverksamheten på förlaget var under avveckling runt 2006, saknades intresse för investeringar i ett nytt system. Lexikograferna gick därför över till att skriva artiklar i Microsoft Word, där de använde sig av ett antal olika mallar.

Sammanfattningsvis kan den situation som råder vad gäller ordboksredigeringssystem på svenska förlag betraktas som beklämmande med tanke på de hjälpmedel som trots allt finns att tillgå. I hög grad är det ekonomin som styr, och ordböcker torde sälja för dåligt för att förlagen ska satsa på ordentliga system. Även om en viss ordbok säljer bra, är det ingen garanti för att bra system införskaffas eller utarbetas. Om ett förlag trots allt satsar på ordboksverksamhet verkar förlaget, som Atkins/Rundell (2008:114) konstaterar, hellre vilja köpa in ett kommersiellt redigeringssystem än utveckla ett eget.

3.2. Ordboksredigeringssystem vid svenska lexikografiska institutioner

En mycket viktig institution när det gäller framställning av ordböcker i Sverige är givetvis redaktionen för *Svenska Akademiens ordbok* (SAOB). Som alla vet har arbetet med ordboken pågått länge, och de tekniska förutsättningarna är idag helt andra än de var då första häftet av SAOB kom ut 1893. Mycket arbete har också lagts ner på att utarbeta ett system för framställning av detta mycket komplexa verk. För närvarande förbereds artiklarna i ett

skräddarsytt redigeringssystem utvecklat av Språkdata vid Institutionen för svenska språket i Göteborg. Övrigt arbete sker i ordbehandlingsprogram – redaktionen går nu över från FrameMaker till Microsoft Word. Källorna behandlas dock i FileMaker och i specialanpassade program skrivna i Java.

En annan svensk lexikografisk institution är Dialektavdelningen vid Institutet för språk och folkminnen i Uppsala, där ett svenskt dialektlexikon samt *Ordbok över folkmålen i Övre Dalarna* (OÖD) sammanställs. Vid arbetet med det svenska dialektlexikonet används en version av Cronoma, utvecklad ur samma redigeringssystem som används vid arbetet med *Lexin Svenska ord* (se avsnitt 4 nedan). *Ordbok över folkmålen i Övre Dalarna* utarbetas däremot i Microsoft Word.

Den tredje lexikografiska institutionen är Lexikaliska institutet vid Göteborgs universitet, där vi båda är verksamma. Vi arbetar just nu, förutom med *Lexin Svenska ord*, i huvudsak med *Svenska Akademiens ordlista* (SAOL) och *Svensk ordbok utgiven av Svenska Akademien* (SO). Vid arbetet med SAOL och SO används två snarlika system baserade på relationsdatabaser implementerade i Ingres och med gränssnitt utvecklade i den grafiska utvecklingsmiljön Open Road från Ingres Corporation. De två systemen är skräddarsydda för de specifika behov som kommit fram under arbetet med dessa verk. Dessa system har också utvecklats av Christian Sjögreen.

Generellt kan man således säga att de lexikografiska institutionerna i Sverige har satsat på att utveckla egna system. Några rent kommersiella system används hitintills inte.

3.3. Ordboksredigeringssystem i övriga Norden: några exempel

Det kan också vara av intresse med några sidoblickar på system som är i bruk vid andra lexikografiska instanser i Norden. Det Danske Sprog- og Litteraturselskab, som står bakom DDO, ODS

och en kommande svensk-dansk ordbok, har valt att satsa på det redan nämnda kommersiella redigeringssystemet iLEX. Tidigare användes GestorLEX.

Inom projektet Norsk Ordbok 2014 arbetar man istället med ett skräddarsytt system, som har presenterats i olika lexikografiska sammanhang (se t.ex. Bakken 2005, Grønvik 2005). Detta system är frukten av ett nära samarbete mellan redaktionen och Eining for digital dokumentasjon vid Universitetet i Oslo.

Inom det samnordiska projektet ISLEX arbetar man med att utveckla en isländsk online-ordbok med isländska som källspråk och danska, norska och svenska som målspråk. Det system som används där är bl.a. beskrivet i Sigurðardóttir et al. (2008). Också där har man satsat på ett skräddarsytt redigeringssystem som utvecklats av och inom projektet.

Utanför Norden används givetvis en rad olika ordboksredigeringssystem. Ett system, vars förtjänster lyfts fram av bl.a. Atkins/Rundell (2008:114), är Dictionary Production System (DPS), som utvecklats av det franska företaget Ingénierie Diffusion Multimédia (IDM). DPS används för olika typer av ordböcker och för ordböcker på många olika språk, men också vid framställning av biografiska lexikon och encyklopedier (se vidare DPS). Ett annat system som återkommer i europeisk lexikografisk litteratur är TshwaneLex dictionary compilation software (även kallat TLex), som utvecklats av TshwaneDJe. Detta belgisk-sydafrikanska system kan – enligt hemsidan – användas för alla typer av ordböcker och ordböcker på alla språk (se vidare Tshwanelex; de Schryver/De Pauw 2007; Joffe, MacLeod/de Schryver 2008). Systemet är sålunda mycket generellt. Ridings (2003:204) påpekar dock att nästan varje ordboksprojekt är unikt. Ett mycket generellt redigeringssystem kan ge lexikografer stora friheter i sin strävan efter att tillgodose så många behov som möjligt. Men för mycket frihet i ett skede av ordboksarbetet kan resultera i mer arbete i ett annat skede, t.ex. vid korrekturläsningen.

I det följande kommer vi att presentera Cronoma, som alltså används vid utarbetandet av *Lexin Svenska ord*.

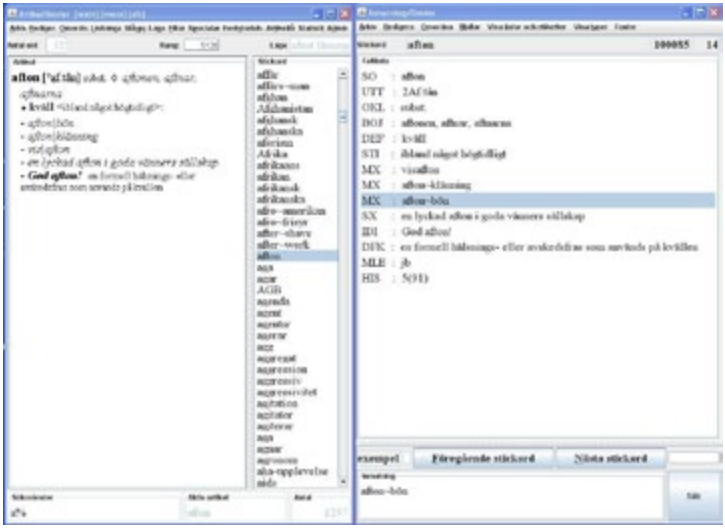
4. Cronoma

Sedan 2008 arbetar vi inom Lexikaliska institutet, på uppdrag av Språkrådet, med att förbättra den svenska orddatabasen i Lexin-projektet (se Gellerstam 1999 för en beskrivning av Lexin-projektet och Pálfi/ Tarp 2009 för en diskussion om innehållet i Lexin-ordböckerna). Arbetet går dels ut på att lägga till nya lemman, dels revidera befintlig information i artiklarna. Det pågående uppdateringsarbetet beskrivs närmare i Hult et al. (2010). För att läsaren ska få en uppfattning om hur Cronoma[2] fungerar återges det användargränssnitt som redaktionen huvudsakligen arbetar med i figur 1. Som utgångspunkt för vår presentation har vi valt artikeln *afton*, som den ser ut i dagsläget.

När lexikografen loggat in i Cronoma möts han eller hon av två fönster, ett artikelfönster (som utgör den vänstra delen av figur 1) och ett inmatningsfönster (som utgör den högra delen). Dessa två från varandra fristående fönster kan dock placeras och utformas som lexikografen önskar på skärmen. Cronoma ger också lexikografen förhållandevis stora friheter vad gäller andra delar av användargränssnittet. Den enskilde lexikografen kan t.ex. välja vilket typsnitt han eller hon föredrar att arbeta med.

Längst ner i vänstra hörnet av artikelfönstret återges den söksträng som skrivits in. I detta fall har vi sökt fram alla lemman

2 Cronoma utvecklades ursprungligen av Christian Sjögreen och Daniel Ridings, Lexilogik AB, för arbetet med *Norstedts ryska ordbok* (2006). En redaktionsmanual skrevs av Sofia Johansson. Cronoma har sedan vidareutvecklats i några olika versioner av Christian Sjögreen. Cronoma är skrivet i Java och kan användas i valfri databashanterare. Det är testat med MySQL, Microsoft Access och Microsoft SQL Server.



Figur 1: Användargränssnitt för Cronoma (artikelfönster och inmatningsfönster)

i orddatabasen som börjar på bokstaven *a* och resultatet av sökningen syns i stickordslistan till höger. I figur 1 återges en del av resultatet av denna lista, närmare ett 30-tal lemman från substantivet *affär* till substantivet *aids*. När lexikografen placerar markören på ett av lemmana i listan och högerklickar framgår det vem i projektet som sist arbetade med artikeln, och när detta skedde. Men det finns också möjlighet att göra mer generella sökningar för att t.ex. få veta hur många artiklar en viss medarbetare arbetat med och vilka dessa är (se vidare Grønvik/Tvedt 2006:148–149 om den här typen av uppgifter som bl.a. kan intressera projektledningen).

I inmatningsfönstret till höger i figur 1 framgår det hur i detta fall artikeln *afton* är uppbyggd. Varje informationskategori i artikeln har en kod: SO = lemma, UTT = uttal, OKL = ordklass, BOJ = böjningsangivelse etc. De olika koderna lär man sig snabbt när man arbetar med systemet, men om man blir osäker finns det en rullgardinsmeny där samtliga koder finns listade med förklaringar.

Om lexikografen skriver in en kod som inte finns med bland de på förhand godkända, uppmärksammas han eller hon på detta genom ett varningsfönster (se vidare Ridings 2003:207 om varningsfönster). För att säkra konsekvensen i angivelserna finns det inom systemet en lista över de ordklassbeteckningar som är godkända av projektledningen. Liknande fasta listor, som också underlättar arbetet för lexikograferna, finns bl.a. för godkända förkortningar. För övrigt kan mycket av arbetet inom systemet ske med hjälp av kortkommandon, vilket gör det praktiska ordboksarbetet mycket smidigt. Dessutom leder det till variation jämfört med det annars vanliga arbetet med datormusen.

Av innehållet i den vänstra delen av artikelfönstret kan man bilda sig en uppfattning om hur artikeln *afton* kommer att se ut. Det faktum att lexikografen har möjlighet att se artikelns innehåll i olika format är enligt bl.a. Rasmussen (2006:355) positivt eftersom olika visningslägen kan tilltala olika användare.

Hittills har vi diskuterat en befintlig artikel i *Lexin Svenska ord*. Vid framställning av nya artiklar väljer lexikografen en mall för den typ av lexikal enhet som det är fråga om (se vidare Svensén 2004:499 om mallar). Dessa mallar har medarbetarna i Lexin-projektet själva skapat – och de kan på ett enkelt sätt revideras av lexikografen om det finns behov av detta. Precis som i de allra flesta redigeringssystem har lexikografen i Cronoma också stora möjligheter att själv söka fram olika delmängder i databasen. Detta är givetvis värdefullt om man framför allt arbetar med t.ex. en viss informationskategori (se vidare Atkins/Rundell 2008:116 om fler viktiga sökmöjligheter som också finns i Cronoma).

5. Slutord

I den här artikeln har vi kortfattat redogjort för några av de många krav som ställs på ett ordboksredigeringssystem av idag. Vidare

har vi redogjort för en mindre undersökning av de system som används för ordboksredigering i Sverige, men också utanför landets gränser, främst i Norden. Trots att ingen inom ordboksbranschen förnekar att ordboksredigeringssystem är viktiga, för systemen en tynande tillvaro på svenska (ordboks)förlag. Inom lexikografiska institutioner tenderar man att ha skräddarsydda system; ännu har inget kommersiellt system fått riktigt fäste i Sverige (jfr t.ex. iLEX i Danmark).

Artikeln ägnas också åt ett visst ordboksredigeringssystem, närmare bestämt Cronoma, som används vid uppdateringen av den svenska Lexin-orddatabasen. Cronoma är ett ordboksredigeringssystem som, enligt vår erfarenhet, passar utmärkt för en ordbok som *Lexin Svenska ord* vilken har en förhållandevis enkel struktur. Dessutom är systemet förhållandevis enkelt att lära sig och arbeta med tack vare bl.a. olika varningsfönster. I Cronoma har också lexikografen stora möjligheter att söka fram delmaterial. En version av detta system kommer inom en relativ snar framtid att bli fritt tillgänglig.

Det beskrivna systemet kan givetvis bli ännu bättre. Atkins/Rundell (2008:116) diskuterar t.ex. möjligheten att ha en stavningskontroll som fungerar i realtid vilket minimerar risken för skrivfel i samband med ordboksarbetet. En annan möjlighet är att redigeringssystemet ger lexikografen stöd att hålla sig till en särskild definitionsvokabulär, något som vi strävar efter inom Lexinprojektet (se vidare Hult et al. 2010). I nuläget finns inte dessa hjälpmedel, men om behov finns skulle de dock kunna arbetas fram. Vår erfarenhet är att ett nära samarbete mellan lexikografer och systemutvecklare är en mycket viktig förutsättning för att ett riktigt bra ordboksystem ska kunna utvecklas och vidmakthållas.

Litteratur

Ordböcker

Bonniers 1980 = Sten Malmström/Iréne Györki: *Bonniers svenska ordbok: vad orden betyder, hur de stavas, hur de uttalas, hur de används, hur de böjs*. 1 uppl. Stockholm: Albert Bonniers förlag.

DDO = *Den Danske Ordbog* 1–6. København: Det Danske Sprog- og Litteraturselskab/Gyldendal 2003–2005.

Lexin Svenska ord 1992: *Lexin Svenska ord – med uttal och förklaringar*. 2 uppl. Stockholm: Norstedts.

Norsk Ordbok = *Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet.* Oslo: Det Norske Samlaget 1966–.

Norstedts ryska ordbok: rysk–svensk, svensk–rysk. 1 uppl. Stockholm: Norstedts 2006.

ODS = *Ordbog over det danske Sprog* 1–28, supplement 1–5. København: Gyldendalske Boghandel 1919–1956, 1992–2005.

OÖD = Levander, Lars/Stig Björklund: *Ordbok över folkmålen i Övre Dalarna*. Uppsala. Skrifter utgivna genom Landsmåls- och folkminnesarkivet i Uppsala [numera genom Institutet för språk och folkminnen] D:1. 1961–.

SAOB = *Ordbok över svenska språket utgiven av Svenska Akademien* 1–. Lund: Gleerups 1898–.

SAOL = *Svenska Akademiens ordlista*. 13 uppl. Stockholm: Norstedts 2006.

SO = *Svensk ordbok utgiven av Svenska Akademien* 2009. Stockholm: Norstedts.

Övrig litteratur

Atkins, B. T. Sue/Michael Rundell 2008: *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bakken, Kristin 2005: Norsk ordbok 2014 – rammer og utfordringar. I: Ruth Vatvedt Fjeld/Dagfinn Worren (red.): *Nordiske studiar i leksikografi 7. Rapport frå Konferanse om leksikografi i Norden Volda 20.–24. mai 2003.* Oslo, 29–35.

de Schryver, Gilles-Maurice/Guy De Pauw 2007: Dictionary Writing Systems (DWS) + Corpus Query Package (CQP): The Case of TshwaneLex. I: *Lexikos* 17 [AFRILEX-series 17:2007], 226–246.

Gellerstam, Martin 1999: LEXIN – lexikon för invandrare. *LexicoNordica* 6, 3–17.

Grønvik, Oddrun 2005: Redigeringsprogrammet for Norsk Ordbok 2014. I skjeringspunktet mellom menneskeleg skjønn og automatisering. I: Ruth Vatvedt Fjeld/Dagfinn Worren (red.): *Nordiske studiar i leksikografi 7. Rapport frå Konferanse om leksikografi i Norden. Volda 20.–24. mai 2003.* Oslo, 157–165.

Grønvik, Oddrun/Lars Jørgen Tvedt 2006: Norsk Ordbok 2014 – presentasjon av eit komplekst leksikografisk verktøy. I: Henrik Lorentzen/Lars Trap-Jensen (red.): *Nordiske studier i leksikografi 8. Rapport fra konference om leksikografi i Norden, Sønderborg 24–28 maj 2005.* København, 143–150.

Hult, Ann-Kristin/Sven-Göran Malmgren/Emma Sköldberg 2010: Lexin – a report from a recycling lexicographic project in the North. I: Anne Dykstra/Tanneke Schoonheim (eds.): *Proceedings of the 14th EURALEX International Congress*, Leeuwarden, 6–10 July 2010. Ljouwert, 800–809.

Joffe, David/Malcolm MacLeod/Gilles-Maurice de Schryver 2008: Software Demonstration: The TshwaneLex Electronic Dictionary System. I: Elisenda Bernal/Janet DeCesaris (eds.): *Proceedings of the XIII Euralex International Congress, Barcelona 15–19 July 2008.* Barcelona, 421–423. [CD-ROM]

Pálfi, Loránd-Levente/Sven Tarp 2009: Lernerlexikographie in Skandinavien – Entwicklung, Kritik und Vorschläge. I: *Lexicographica* 25, 135–154.

Rasmussen, Marie Bilde 2006: Nyt redigeringssystem – overvejelser og valg. I: Henrik Lorentzen/Lars Trap-Jensen (red.): *Nordiske studier i leksikografi 8. Rapport fra konference om leksikografi i Norden, Sønderborg 24–28 maj 2005*. København, 347–357.

Ridings, Daniel 2003: Lexicographic workbench: A case history. I: Piet van Sterkenburg (ed.): *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins, 204–214.

Svensén, Bo 2004: *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik.* 2 upplagan. Stockholm: Norstedts Akademiska förlag.

Sigurðardóttir, Aldis/Anna Hannesdóttir/Håkan Jansson/Halldóra Jónsdóttir/Lars Trap-Jensen/Þórdís Úlfarsdóttir 2008: ISLEX – An Icelandic-Scandinavian Multilingual Online Dictionary. I: Elisenda Bernal/Janet DeCesaris (eds.): *Proceedings of the XIII Euralex International Congress, Barcelona 15–19 July 2008.* Barcelona, 779–789. [CD-ROM]

Internethänvisningar

iLEX = http://www.emp.dk/ilexweb/ (maj 2010).

Lexique Pro = http://www.lexiquepro.com/ (maj 2010).

DPS = http://www.idm.fr/products/dictionary_writing_system/27/ (maj 2010).

Tshwanelex = http://tshwanedje.com/ (maj 2010)

Anmärkning: Emma Sköldberg har skrivit sin del av artikeln inom ramen för sin forskartjänst som finansieras av Svenska Akademien.

Christian Sjögreen
systemutvecklare
Lexikaliska institutet
Institutionen för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
christian.sjogreen@svenska.gu.se

Emma Sköldberg
fil. dr, forskare
Lexikaliska institutet
Institutionen för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
emma.skoldberg@svenska.gu.se

Felles leksikalske ressursar for språkteknologi og leksikografi

Trond Trosterud

Lexicography and language technology have traditionally been seen as two distinct disciplines. In addition to pointing at the common need to manage huge lemma lists in one way or another, the present paper argues for conducting lexicographical and language technology projects in an integrated way, with the same lemma list for both dictionary and grammatical analyser/generator. This is especially important for small language communities, which lack the resources to uphold several parallel lexical databases.

1. Innleiing

La oss byrje med å sjå verda frå språkteknologen og leksikografen sin ståstad. Språkteknologen kan analysere laupande tekst, og vil gje analysen kunnskap om orda. Leksikografen vil vite kva eigenskapar lemmaforrådet har, og presentere dette for ordboksbrukaren, for eitt og eitt lemma om gangen. Som regel er det lite kontakt desse miljøa imellom. For små språksamfunn er det problematisk. Større språksamfunn har ressursar til å vedlikehalde store grammatiske og leksikalske ressursar i fleire miljø, små språksamfunn kan ikkje det.

Ein integrasjon mellom leksikografi og språkteknologi kan t.d. innebere at leksikonet i den morfologiske analysatoren blir erstatta med ordboka si lemmaliste, slik at den informasjonen som analysatoren treng, også blir lagt til basen som er grunnlaget for ordboka. Den morfologiske analysatoren treng lemma, stamme og informasjon om kontinuasjonsleksikon (bøyingsklasse). Den syn-

taktiske analysatoren treng informasjon om syntaktiske funksjonar og valens, og informasjon om semantiske roller.

Denne artikkelen vil argumentere for å integrere delar av det leksikografiske og språkteknologiske arbeidet, der same lemmaliste inneheld både leksikografisk og språkteknologisk informasjon. I tillegg til å spare arbeid til vedlikehald av store lemmalister, vil det også gje meir avanserte ordbøker og analyseprogram.

2. Leksikografi og språkteknologi som separate prosjekt

Det er lett å innsjå det problematiske ved å vedlikehalde fleire parallelle lemmalister, særleg for språksamfunn med avgrensa ressursar. Korfor er den vanlegaste arbeidsmåten likevel at leksikografane og språkteknologane arbeider kvar for seg?

Det viktigaste argumentet for å operere med leksikalske basar som berre tener eitt formål kvar, er at sjølve basen, eller kjeldekoden, blir enklare. Viss kvart lemma får knytt til seg berre ein handfull typar av data, kan basen bli lagra som ei kommaseparert liste (t.d. i form av eit rekneark) med berre dei opplysningane som trengst til eitt spesifikt prosjekt. Ein slik framgangsmåte kan vere kostnadseffektivt for akkurat det relevante formålet. Det er også lettare å finne hyllevareprogramvare til eitt veldefinert formål. Ein ordboksbase til ei ordbok med eitt formål er lettare å tilpasse til eit ordboksredigeringsprogram enn ein base som skal brukast til mange formål.

Argument mot å la same leksikalske base ha mange formål, er at det blir stort og komplisert. Prosjekta risikerer å velte på lange diskusjonar om leksikonstruktur; innafor språkteknologien er dette er ei velprøvd tue som har velta mange lass. Slike fleirbruksbasar stiller også større krav til datakompetansen til brukarane. Det er med andre ord problematisk å lage fleirbruksbasar, og det er enklare på kort sikt å halde basar for ulike føremål kvar for seg.

I den leksikografiske kvardagen ser vi at akademiske ordboksprosjekt ofte er store prosjekt med lange tradisjonar, og ikkje knytte til språkteknologi. Kommersielle ordboksprosjekt på si side opererer under strenge økonomiske vilkår. Einforfattarprosjekt vil ofte heller ikkje ha kunnskap til eit vidare prosjekt. Alt dette talar for å oppretthalde status quo, og la ulike prosjekt arbeide på kvar sine leksika, heller enn å la same leksikon tene både leksikografiske og språkteknologiske føremål.

I den språkteknologiske kvardagen er akademiske språkteknologiske prosjekt sjeldan innretta på å analysere laupande tekst på ein måte som er robust nok til faktisk å gje gode resultat for autentiske tekstar. Prosjekt for semantisk annotering er som regel ikkje integrert med andre språkteknologiske eller leksikografiske komponentar. Kommersielle språkteknologiske prosjekt opererer, på same måten som kommersielle ordboksprosjekt, under strenge økonomiske vilkår. Maskinomsetjing var tidlegare eit viktig felt for datamaskinell leksikografi, ved utarbeiding av avanserte transferleksika, dvs. ordbøker til bruk for maskinomsetjingsprogramma. Frå og med midten av 1990-talet kom det eit paradigmeskifte innafor maskinomsetjingsfeltet, fokus vart flytt frå grammatisk til statistisk basert maskinomsetjing, og maskinomsetjingsfeltet har seinare vore lite interessert i transferleksika med semantisk informasjon (om valens o.l.). Alt dette gjev dårlege vilkår for kontakt mellom leksikografi og språkteknologi. Også innafor språkteknologien er det altså fleire faktorar som dreg i retning av status quo.

3. Integrasjon mellom leksikografi og språkteknologi

Ei felles lemmaliste og ein felles leksikalsk database for både ordbøker og språkteknologiske applikasjonar er lettare å vedlikehalde enn å ha parallelle leksikalske ressursar. Men det er også andre fordelar ved å ha ein felles database.

Der lemmaforrådet er direkte knytt til ein morfologisk analysator, har ordboksbrukaren tilgang ikkje berre til oppslagsforma, men til alle ordformene i bøyingsparadigmet. Dette er spesielt viktig for elektroniske resepsjonsordbøker. Ein morfologisk analysator knytt til ordboka gjer dette mogleg. Dei kommersielt interessante språka er for morfologifattige til å tvinge fram integrerte grammatiske analysar i dei elektroniske ordbøkene. Det vesle som finst av morfologi i t.d. engelsk og tysk kan leggjast til manuelt eller ved enkle prosessar, og løysingar for morfologirike språk blir ikkje utvikla.

Ved å integrere leksikon med morfologisk analyse vil den grammatiske analysatoren ha tilgang til informasjon frå ordboka. Viktig her er valens og semantikk. For å få ein vellukka grammatisk analyse er det viktig å ha t.d. valensinformasjon for verb. Ein NP i oblik kasus kan bli analysert som oblikt objekt eller laust adverbial, alt etter eigenskapane til verbet. Omvendt vil også aktiv bruk av den grammatiske informasjonen i ordboka setje denne informasjonen på prøve: Viss t.d. valensramma seier at eit visst verb ikkje skal ta akkusativobjekt, men vi ved å kombinere leksikon med ein analysator og ein tekst finn setningar der verbet faktisk tar akkusativobjekt likevel, blir vi nøydde til å revidere ordboka på det punktet.

4. Korleis integrere

Eit eksempel på integrerte løysingar er arbeidet med leksikografi og språkteknologi for komi som blir gjort i Helsingfors og Tromsø. Komi er eit finskugrisk språk i Nordvest-Russland. Arbeidet har ikkje kome langt, men integreringa av leksikografi og språkteknologi er på plass. Bakgrunnen for arbeidet er Jack Rueters *Komi–English–Finnish Dictionary* (KEFD) og analysatoren for komi (KOMFST). Poenget med eksempelet er altså verken ordboka eller

analyseprogrammet, men måten dei er integrert med kvarandre på.

Figur 1 viser ein tilfeldig ordboksartikkel for ordboka komi–engelsk/finsk, verbet аддзödчыны *'addzödtsjyny'* "å møte". Relevant her er, i tillegg til lemma, også stamme (<stem>), ordklasse (<pos>), og kontinuasjonsleksikon (dvs. ein peikar til kor i automaten vi skal gå for å lese neste tilstand (<contlex>)).

```
<entry>
   <lemma>аддзödчыны</lemma>
   <stem>аддзödчы</stem>
   <contlex>Verb3</contlex>
   <pos>V</pos>
   <article>
      <sem>RECIP</sem>
      <syn>SVP/-кöд</syn>
      <eng>
         <choice>
            <variant>meet</variant>
         </choice>
      </eng>
      <fin>
         <choice>
            <variant>tavata</variant>
         </choice>
      </fin>
   </article>
</entry>
```

Figur 1: Komi-leksikon i XML-format

Ordboka er integrert med kjeldekoden til ein endeleg tilstandsautomat.[1] Kvar ordklasse genererer ei leksikonfil i automaten, som

1 Ein endeleg tilstandsautomat (*finite-state automaton*) er ein modell som inneheld eit endeleg sett av *tilstandar* og overgangar frå tilstand til tilstand, der *tilstand* er definert som eit unikt sett av informasjon. Ein automat for ordet *båt* vil bestå av fire tilstandar: ein starttilstand, ein tilstand for *b*, ein for *bå*, og ein sluttilstand for *båt*. Ein spesiell type tilstandsautomatar er transdusarar, som ikkje berre les symbol, men også endrar dei, slik at strengen *båtane* kan bli endra til strengen *båt +N +Sg +Def* og omvendt. For ei innføring i bruk av endelege tilstandsautomatar i språkteknologi, sjå Beesley og Karttunen (2003).

par av lemma (til venstre for kolon) og stamme (til høgre for kolon), som vist i Figur 2. Frå denne representasjonen peiker kontinuasjonsleksikonet (her: Verb3) vidare i automaten. For å gjere leksikonet lettare å lese er det også teke med ei omsetjing, men berre som kommentarfelt til internt bruk.

```
абутӧмасьны:абутӧмась Verb2  "feign poverty" ;
адавны:адал Verb1  "gobble up" ;
аддзавны:аддзал Verb1  "find" ;
аддзӧдчыны:аддзӧдчы Verb3  "meet" ;
аддзӧдлыны:аддзӧдлы Verb3  "see" ;
аддзыны:аддзы Verb3  "see" ;
аддзывны:аддзыл Verb3  "experience" ;
аддзысьны:аддзысь Verb2  "be found" ;
```

Figur 2: Leksikonfil i automaten for komi

I verbmorfologifila i Figur 3 får verbet infinitivsending, og det blir sendt vidare til leksikonet Finiteforms, der det får tempus, før det blir sendt vidare til leksikona for personendingar. Resultatet er ein morfologisk generator/analysator for komi.

```
LEXICON Verb3 ! ярмыны. Ending in -ыны.
+V:  VerbConj ;
+V+Inf:%>ыны K ;

LEXICON VerbConj
 Finiteforms ;
 Non-finiteforms ;

LEXICON Finiteforms  ! Gives linking vowels for 3 tenses
+Ind+Fut:%>а PresPret1 ; ! 1,2  Future
+Ind+Prs:%>а PresPret1 ; ! 1,2  Present
+Ind+Prt1:%>і PresPret1 ; ! 1,2  Preterite1. і и variation

LEXICON PresPret1  ! First and second person
+Sg1:%> K ;
+Sg2:%>н K ;
+Pl1:%>м K ;
+Pl1:%>мӧ K ;
+Pl1:%>мӧй K ;
+Pl2:%>ныд K ;
+Pl2:%>д K ;
```

Figur 3: Morfologisk generator/analysator for Komi (utdrag)

Med utgangspunkt i denne automaten er det mogleg å analysere tekst på komi. Figur 4 viser ein analyse av presentasjonsteksten på framsida til komiutgåva av Wikipedia. Komi-analysatoren er uferdig, og nye lånord og namn er ikkje med.

Figur 4: Analysert tekst frå komi Wikipedia, med manglande dekning for lånord som субдоменсö 'subdomene' og Википедия 'Vikipedija'

I tillegg til å danne kjernen i ein morfologisk analysator, kan den leksikalske databasen sjølvsagt også brukast til å lage ordbøker, t.d. e-ordbøker, som i Figur 5. Her er på ny ordet for å møte, *аддзöдчыны*, og same lemmaartikkel er no presentert i eit ordboksgrensesnitt. Ordforma er ei infinitivsform, jf. endinga *-ыны*, men frå denne ordboka sitt synspunkt er ordforma eit uanalysert heile.

Figur 5: Databasen til automaten fungerer også som input for ei komi e-ordbok

Så lenge den morfologiske analysatoren er kompatibel med informasjonen om bøyingsklasse og stamme som blir oppgjeven i ordboka, er det mogleg å bruke same morfologiske analysator til fleire ordbøker. Eit døme på det er færøysk.

Fróðskaparsetur Føroya har gjeve ut ei svært god einspråkleg ordbok, *Føroysk orðabók* (FO). Dei har deretter gjort lemmaliste og bøyingskode frå ordboka tilgjengeleg for språkteknologiske føremål. Denne lista utgjer leksikon for analysatoren for færøysk utarbeidd ved Universitetet i Tromsø (FAOFST, Trosterud 2009). Analysatoren FAOFST bruker nøyaktig same bøyingskoder som FO, og kan med hjelp av denne informasjonen generere alle paradigma for alle oppslagsorda i ordboka. Dette inneber at det også er mogleg å kombinere nye lemmalister, t.d. nye, reviderte utgåver av ordboka, eller den færøyske delen av eventuelle tospråklege ordbøker, med analysatoren, så lenge desse andre ordbøkene også bruker FO sitt bøyingskodesystem. Eit døme på ein slik situasjon har vi for finsk, der *Suomen kielen perussanakirja* (SKP) sitt bøyingsklassesystem har vorte ein de facto standard for fleirspråklege ordbøker, t.d. til marisk og norsk (jf. Trosterud 2003:13ff for ei drøfting). Det same bøyingsklassesystemet ligg til grunn for ein ope tilgjengeleg morfologisk analysator for finsk (OMORFI, Pirinen 2008).

5. Bruk av språkteknologi i leksikografien

Ein føresetnad for at det skal vere mogleg å gjere bruk av automatiske analysemodellar i leksikografien, er at dei leksikografiske databasane er maskinleselege. Dei ulike delane av lemmaartikkelen må formaterast semantisk, og ikkje visuelt, dvs. etter kva dei ulike felta inneheld og ikkje etter korleis dei ser ut. Innhaldet i felta for grammatisk informasjon må også vere utvitydig maskinleseleg. T.d. er det ikkje nok å gje eit fullt bøyingsmønster (med

fleirtalsformer inkludert) for eit abstrakt substantiv som berre opptrer i eintal. For menneskelege lesarar er det mogleg å stole på at dei forstår at substantivet det gjeld, ikkje blir brukt i fleirtal, og dermed heller ikkje har fleirtalsformer, men slike subtile kombinasjonar av direkte formalisert og indirekte implisitt informasjon er for vanskelege for maskinell prosessering. Skal det vere mogleg å representere den informasjonen som potensielt ligg i ordboka, må han bli gjort eksplisitt. Det må, for kvart av leksema utan fleirtalsformer, faktisk bli brukt ein annan kode enn for elles identiske leksem som altså har fleirtalsformer.

Ein måte å bruke språkteknologi i leksikografien er ved utarbeiding av elektroniske resepsjonsordbøker. Som vist i Antonsen m.fl. 2009, vil ei resepsjonsordbok basert berre på oppslagsformene i ei mellomstor ordbok (testen vart utført med lemmaforrådet i *Bokmålsordboka* (BO), SKP og *Sámi-suoma sátnegirji* (SSS)) kjenne att berre 30,5 % av orda i laupande tekst for norsk, 10 % for finsk, og 7,9 % for samisk (jf. Antonsen m.fl. 2009). Ein stor del av dei ordformene som ordbøkene ikkje kjenner att, vil ha regelrett morfologi og dermed vere lett attkjennelege for lesarar, men så lenge ordboka ikkje kjenner dei att, vil det vere umogleg å klikke på desse og få omsetjingar direkte i lesing av tekst på skjerm.

Neste steg for den elektroniske resepsjonsordboka er disambiguering og leksikalsk seleksjon. Med tilgang til heile setninga vil det for ordboka vere mogleg å finne rett ordklasse og rett homonym, og å prioritere mellom ulike tydingar. Det syntaktiske rammeverket som har vist seg robust nok til å analysere laupande tekst, er *føringsgrammatikk* (*Constraint Grammar*, jf. t.d. Karlsson et al 1995). Føringsgrammatikalske analyseprogram tar setningar (eller større einingar) som er morfologisk analyserte som i Figur 4, som input, vel rett morfologisk analyse og legg til syntaktiske funksjonar. I ei setning som *Vi skal finne mat* vil eit sett føringsgrammatiske reglar velje infinitivslesinga for *finne* pga. modalverbet til venstre, og forkaste imperativslesinga av *mat* pga. det transi-

tive verbet til venstre. Korrekt grammatisk analyse vil dermed (for ei norsk–engelsk e-ordbok) unngå omsetjingsframlegg som *Finn* og *feed*, til fordel for *find* og *food.* Same type reglar kan også skilje mellom ulike tydingar av same ord, som i setningane *Toget er i rute* og *Toget hadde overraskande mange deltakarar*, der substantiva til slutt i setningane peiker på omsetjingar som *train* og *march.*

Det å setje saman leksikografi og språkteknologi opnar også nye perspektiv for vurdering av lemmaforrådet generelt, og dermed også for ordbokskritikken. Eit godt døme på det er arbeidet til Hurskainen (2004), ein ordbokskritikk som kviler på eit noko uvanleg forarbeid. For å kunne evaluere fem ulike ordbøker med swahili som L1 har Hurskainen fått tilgjenge til elektroniske versjonar av alle fem, og deretter kopla dei alle til ein morfologisk analysator over swahili. Deretter har han analysert eit balansert tekstkorpus på omtrent 4 millionar ord. På den måten har han vore i stand til å vurdere dei ulike ordbøkene etter to ulike kriterium: I kor stor grad dei dekkjer ordforrådet i korpuset, og i kor stor grad dei inneheld ord som ikkje finst i korpuset.

Det første kriteriet tilsvarer ein vanleg framgangsmåte ved ordbokskritikk: Å undersøke i kor stor grad ordforrådet i eit visst tekstkorpus er dekt av ordboka. Skilnaden er at der vi normalt er nøydde til å ta stikkprøver, kan Hurskainen faktisk analysere heile tekstsamlingar. Med det andre kriteriet undersøker Hurskainen i kor stor grad ordboka inneheld såkalla ordboksord, ord som berre finst i ordbøker, men aldri i aktuell språkbruk. Viss vi finn ord som verkeleg aldri opptrer utanfor ordboka, må dei sjølvsagt ut. Men manglande treff i korpuset for ord som er ein del av det mentale leksikonet vårt, kan også bli brukt til å evaluere korpusmaterialet vårt. Dersom vi t.d. kan vise at det i ordboka finst ord som ikkje er belagde i store korpus på fleire hundre millionar ord (eller ikkje på sjølve Internett), og det viser seg at desse orda faktisk inngår i språkkunnskapen til morsmålsinformantar, kan denne metoden vise systematiske slagsider ved korpussamlingane. Arvi Hurskai-

nen sitt poeng er likevel å vise kva av ordbøkene hans som på den mest økonomiske måten (med færrast moglege lemma) er i stand til å dekkje størst mogleg del av testkorpuset.

Føresetnaden for å kunne gjere nytte av språkteknologi i leksikografisk basert programvare er at det leksikografiske grunnlagsmaterialet i seg sjølv er strukturert på ein maskinleseleg måte.

6. Bruk av leksikografisk informasjon i språkteknologien

For språkteknologien er lemmalista leksikografanes viktigaste bidrag. Utan lemmaliste blir det ingen analysator. Deretter gjev ordboka bøyingsklasseinformasjonen for kvart einskilt lemma, og ein god ordboksgrammatikk knytt til ordboka og bøyingsklasseinformasjonen vil vere betre enn ein allmenn grammatikk, i og med at ordboksgrammatikken er laga for å gje informasjon om bøyinga til kvart einskilt lemma.

Mange ordbøker har valensinformasjon, enten i form av eksempelsetningar eller i form av kodar for ulike valensklasser. Brukarar og maskiner har ulike preferansar; der maskinene må ha eintydige kategoriar, må brukarane ha døme eller ordformer det er lett å hugse tydinga av. Databasen kan sjølvsagt innehalde begge delar. Viss denne informasjonen blir formalisert, kan han også bidra til språkteknologisk bruk, som t.d. grammatisk disambiguering og syntaktisk analyse.

7. Konklusjon

Leksikografien er framleis langt frå å utnytte potensialet som ligg i å integrere ordbøker og språkteknologi. Dette er viktig både for utvikling av leksikografien, for effektiv evaluering av ordbøker, og

for brukaren, som får tilgang til nye bruksområde for ordbøkene. Språkteknologien på si side har no robuste analysatorar og kan utnytte den semantiske kunnskapen leksikografane har skaffa fram.

På denne måten vil den integrerte leksikografien og språkteknologien oppfylle grammatikaren sin draum om å modellere menneskets språkkunnskap: kunnskap om alle orda – og om korleis dei kan bli brukt.

Litteratur

Ordbøker

BO = Boye Wangensteen (red.) 2005: *Bokmålsordboka. Definisjons- og rettskrivningsordbok*. 3. utgave. Oslo: Kunnskapsforlaget.

FO = Jóhan Hendrik W. Poulsen, Marjun Simonsen, Jógvan í Lon Jacobsen, Anfinnur Johansen og Zakaris Svabo Hansen 1998: *Føroysk orðabók*. Band 1–2. Tórshavn: Føroya Fróðskaparfelag.

SKP = Risto Haarala m.fl. 2001: *Suomen kielen perussanakirja*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 55. Helsinki.

SSS = Pekka Sammallahti 1989: *Sámi-suoma sátnegirji = Saamelais-suomalainen sanakirja*. Ohcejohka: Jorgaleaddji.

Annan litteratur

Antonsen, Lene/Trond Trosterud/Ciprian-Virgil Gerstenberger/Sjur Nørstebø Moshagen 2009: Ei intelligent ordbok for samisk. I: *LexicoNordica 16*, 271–283.

Beesley, Kenneth R. og Lauri Karttunen 2003: *Finite State Morphology*. Palo Alto, CA: CSLI Publications.

Hurskainen, Arvi 2004: Computational testing of five Swahili dictionaries. I: *20th Scandinavian Conference of Linguistics*. http://www.ling.helsinki.fi/kielitiede/20scl/Hurskainen.pdf

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä og Arto Anttila (red.) 1995: *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Natural Language Processing, No 4. Berlin/New York: Mouton de Gruyter.

Pirinen, Tommi 2008: Suomen kielen äärellistilainen automaattinen morfologinen analyysi avoimen lähdekoodin menetelmin, pro gradu -tutkielma. http://www.helsinki.fi/~tapirine/gradu/Pirinen2008.pdf

Trosterud, Trond 2003: Ordbokskritikk. I: *LexicoNordica 10*, 65–88.

Trosterud, Trond 2009: A constraint grammar for Faroese. I: *NEALT Proceedings Series* 2009; Vol. 8. 1–7. http://dspace.utlib.ee/dspace/handle/10062/14285

Internettreferansar

FAOFST = Trond Trosterud 2009: *Færøysk morfologisk analyseprogram.* http://giellatekno.uit.no/cgi/index.fao.nno.html

KEFD = Jack Rueter/Ciprian Gerstenberger/Trond Trosterud 2010: *Komi–English–Finnish Dictionary.* https://victorio.uit.no/langtech/trunk/kt/kom/src/working_files/

KOMFST = Trond Trosterud/Ciprian Gerstenberger/Jack Rueter 2010: *Komi morfologisk analyseprogram*. http://giellatekno.uit.no/cgi/index.kom.nno.html

Trond Trosterud
førsteamanuensis, ph.d.
Fakultet for humaniora, samfunnsvitenskap og lærerutdanning
Universitetet i Tromsø
NO-9037 Tromsø
trond.trosterud@uit.no

IKKE-TEMATISKE BIDRAG

Bilingvale ordbøger med dansk og ungarsk

Loránd-Levente Pálfi, Erzsébet Stokholm & Sven Tarp

Bilingual lexicography for Danish and Hungarian is a subject that has been and still is very scarcely dealt with in the Nordic Countries. As far as we know, no scholarly literature written in Danish has ever been produced on the topic. The motivation for this paper comes from the fact that a new medium-size Danish–Hungarian dictionary has newly seen the light of day. Instead of writing a review of the book, we have used the opportunity to write this review article, with which we partly want to give a complete overview of bilingual dictionaries for Danish and Hungarian from the beginning up until today, which is from 1945 till 2008, and partly a meta-lexicographic critique and analysis of the most important dictionaries.

1. Bilingval leksikografi

Termen *bilingval ordbog* er på sin vis en sekundær term i metaleksikografien, og det samme er følgelig termen *bilingval leksikografi* (jf. Tarp 2004). Begge termer refererer nemlig til formmæssige træk ved leksikografiske opslagsværker og siger ikke i sig selv noget om, hvad specifikke ordbøger kan bruges til, og hvem der kan bruge dem.

Inden for den leksikografiske funktionslære betragtes ordbøger frem for alt som brugsgenstande, hvis formål ligesom ved alle andre brugsgenstande er at tilfredsstille bestemte menneskelige behov; i leksikografiens tilfælde de specifikke typer af punktuelle informationsbehov, der kan opstå hos specifikke typer af potentielle brugere i specifikke typer af ekstra-leksikografiske situationer (jf. bl.a. Tarp 2006). Det afgørende kendetegn ved et hvilket som helst leksikografisk opslagsværk er derfor, hvilken funktion

eller hvilke funktioner det er konciperet til at varetage. Disse funktioner kan være *kognitive* med henblik på at give brugeren ny viden, eller *kommunikative* med henblik på at løse problemer i forbindelse med produktion, reception og oversættelse af tekster.

Ordbøger konciperet til kommunikative og kognitive formål på modersmålet er i sagens natur monolingvale. Men i forbindelse med fremmedsprog eksisterer der en kompleks relation mellem ordbøgers mono- eller bilingvalitet og deres mulige funktioner. Som eksempel kan nævnes, at en avanceret fremmedsproglig lørner i de fleste tilfælde vil kunne nøjes med en monolingval ordbog på fremmedsproget for at løse sine forskelligartede kommunikative problemer i forbindelse med fremmedsproglig tekstproduktion, hvorimod en lørner på begynderniveau ofte vil have brug for en kombination af en bilingval ordbog fra modersmål til fremmedsprog og en monolingval eller tilmed bilingval ordbog, der tager udgangspunkt i fremmedsproget.

I det hele taget gælder, at brugere, der søger leksikografisk hjælp til at løse problemer i forbindelse med de forskellige former for kommunikation på fremmedsproget, i de fleste tilfælde vil have behov for en kombination af en mono- og bilingval ordbog eller en kombination af to bilingvale ordbøger i hver sin sprogretning. Én bilingval ordbog i en enkelt sprogretning vil derfor sjældent være tilstrækkelig til at løse alle brugerens behov i forbindelse med en bestemt type fremmedsproglig kommunikation. Følgelig er det vanskeligt at foretage en udtømmende bedømmelse af en sådan ordbogs mangler og fortræffeligheder, hvis den ikke indgår i en større "pakke" med en eller flere andre ordbøger. Og samtidig har ingen nok så raffinerede lingvistiske data nogen relevans, hvis de ikke er medtaget med henblik på løsning af den forudsete brugergruppes forudsete problemer i de forudsete situationer (hensigtsløse data kan dog funktionaliseres og dermed blive relevante). Med disse forbehold og begrænsninger vil de hidtidige bilingvale ordbøger med dansk og ungarsk blive undersøgt i det følgende.

2. Fra Világi (1945) til Nielsen et al. (2008)

Antallet af bilingvale ordbøger med dansk og ungarsk løber op i blot fem bogtrykte værker, af hvilke det ene (den dansk–ungarske ordbog ved Nielsen et al.) er udkommet i tre udgaver – 1982, 1993 og 2008 – foruden i et uændret optryk af 1993-udgaven i 1998. I den følgende opstilling anføres ophav, udgivelsestidspunkt, forkortet titel, side- og lemmaantal:

Világi, Zoltán	1945	*Dán–magyar...*	79 pp.	5.000
DBK	1957	*Dansk–ungarnsk* [sic!]...	83 pp.	4.900
Knudsen, Jørgen	1958	*Dán–magyar és...*	145 pp.	7.300
Nielsen, Margit et al.	1982	*Dansk–ungarsk ordbog...*	??? pp.	14.000
Nielsen, Margit et al.	1993	*Dansk–ungarsk ordbog...*	638 pp.	37.000
Nielsen, Margit et al.	1997	*Magyar–dán szótár...*	848 pp.	51.000
Nielsen, Margit et al.	2008	*Dansk–ungarsk ordbog...*	533 pp.	44.500

Mens Nielsen et al. (1982, 1993 og 2008) er dansk–ungarsk-ordbøger og Nielsen et al. (1997) en ungarsk–dansk-ordbog, indeholder Világi (1945), DBK (1957) og Knudsen (1958) to separate opslagsdele hver og er således dansk↔ungarsk-ordbøger.

2.1. Világi (1945)

I de sidste uger af Anden Verdenskrig blev 12.000 ungarske soldater tvangsudskrevet og sendt i tysk krigstjeneste til Danmark. I denne sammenhæng opstod pludselig et behov for en ordbog med ungarsk og dansk, og det blev officeren Zoltán Világi, der udførte arbejdet.[1] Få dage efter at bogen blev trykt, havde den tyske

1 Det var ikke et ufarligt arbejde: Da Gestapo fik kendskab til værket, udstedtes en arrestordre mod Világi; den blev dog ikke effektueret pga. kapitulationen (Sørensen 2005:71–72).

militærmagt imidlertid kapituleret. Ifølge Sørensen (2005:71) blev bogen trykt i et oplag på 650 eksemplarer.

I det følgende ses et udsnit gengivet så tro mod forlægget som muligt:

> **E**cet [sic!], Eddike
> [...]
> egyæbkænt [sic!], forøvrigt
> egyedül, alene – ene
> egyedülállò [sic!], enestaaende

Ud over de mest nødvendige præliminærsider[2] findes der ingen omtekst. I forordet meddeles, at de særlige ungarske vokaler (Á, á, É, é, Í, í, Ó, ó, Ö, ö, Ő, ő, Ú, ú, Ü, ü, Ű, ű) ikke i alle tilfælde er gengivet korrekt, hvorfor der undertiden bruges dansk æ for ungarsk é, dansk å for ungarsk á osv. Hvorfor denne praksis ikke gælder alle tilfælde, og hvorfor der ikke er konsekvens (altid Æ/æ for É/é osv.), forklares ikke. Alfabetiseringen er fra Aa til Ø, danske substantiver skrives med stort begyndelsesbogstav, og der findes ingen kolumnetitel. Ækvivalenter adskilles fra lemmata blot ved et komma og ikke ved typografiske strukturindikatorer; til markering af bogstavskifte bruges dog i det første bogstav af det første ord en anden typografi, fed skrift samt majuskel.

2.2. DBK (1957) og Knudsen (1958)

Tilblivelsen af DBK (1957) og Knudsen (1958) må også forklares på en politisk baggrund: opstanden i Ungarn i oktober/november 1956, som medførte, at ca. 200.000 ungarere flygtede – Danmark tog imod 1.400.

2 Præliminærsider er de sider, som går forud for en bogs hovedtekst, og omfatter traditionelt smudstitelblad, titelblad, kolofon, indholdsfortegnelse, illustrationsfortegnelse, forord og indledning; se evt. *Informationsordbogen* (Dansk Standard, 2002).

Ud over titelbladene (ét til hver opslagsdel) indeholder DBK (1957) ingen omtekst; i hvert fald ikke det eksemplar, vi har lånt på Statsbiblioteket. Alfabetiseringen sker fra Å til Ø. Brugen af stort begyndelsesbogstav i de danske substantiver i begge opslagsdele samt i den levende kolumnetitel er overraskende. I det følgende ses et udsnit fra s. 20 gengivet så tro mod forlægget som muligt:

ialfald, mindeneseiz [sic!]
iboende, bennlakó
Idræt, en, sport
[...]
ifølge, szering [sic!], utár [sic!]
iføre, felhuz [sic!], ruház

Værket fremviser en del påfaldende lemmaformer, f.eks. **ialfald** i stedet for *i al fald*, samt mange trykfejl og ækvivalensfejl – f.eks *mindeneseiz*, *szering* og *utár* (burde have været *mindenesetre*, *szerint* og *után*), **iboende** (kan ikke oversættes med *bennlakó*, som betyder 'person, som bor på et internat eller kollegium', men med *benne rejlő* eller *benne lakozó*), **iføre** (ikke *ruház*, men *felvesz* og *felölt* foruden *felhúz*, som i øvrigt staves med "ú"). Bortset fra sprogretningen er der ingen nævneværdig forskel imellem de to opslagsdele.

Knudsen (1958) fremviser en mere kompleks fordelingsstruktur: Foruden titelblad og opslagsdel indeholder værket en ordbogsgrammatik (på 21 sider) med bl.a. en forkortelsesliste og en liste over uregelmæssige verber. Også i dette værk benyttes en levende kolumnetitel. I det følgende ses et udsnit gengivet så tro mod forlægget som muligt:

```
SOLID komoly, tartós
[...]
SOM mint, amint, úgymint, ahogy. SOM BEKENDT amint
  ismeretes. SOM DU VIL ahogy akarod. SOM OM mintha
SOMME TIDER időnként, olykor
SOMMER/,en nyár. ~FUGL,en lepke
```

```
[...]
SLID/,et erőfeszítés, elhasználódás.
  ~ER elnyű, szorgoskodik
SLIKKER nyal, nyalakodik
```

Artiklerne indeholder (1) lemma sat med kapitæler og i mange tilfælde med inkorporeret angivelse af tryk (f.eks. **SOLID**), (2) for substantivers vedkommende angivelse af køn, samt (3) angivelse af en (eller flere) ækvivalent(er). I mange tilfælde anføres (4) flerordsenheder sat med kapitæler og efterfulgt af en (eller flere) ækvivalent(er) (f.eks. **SOM BEKENDT**, **SOM DU VIL** og **SOM OM**). Ordningen af lemmata er nichealfabetisk (jf. **SOMMER** og **SLID**), og verber angives i tredje person singularis, hvilket er gængs praksis i ungarsk leksikografi, jf. **SLID/[...]~ER** og **SLIKKER**. Værket fremviser en meget simpel intrakomponentiel mediostruktur, idet der kun sjældent henvises inden for samme ordbogsdel (f.eks. fra **NR.** til **NUMMER**). Den ungarsk–danske opslagsdel er strukturelt ikke helt identisk med den dansk–ungarske: den fremviser et meget simpelt interkomponentielt henvisningssystem:

```
CÉKLA rødbede-n,-r (tsz!)
CÉL/ mål-et,-. (szándék) formål-et,-.
  ~OZ sigter 1. ~TALAN formålsløs
[...]
CIPŐ/ (fél-) sko-en,- (tsz!). (füzős,
  magas) støvle-n,-r (tsz) ~FÜZŐ
  snørebånd-et,- (tsz!). ~KANÁL sko-
  horn-et,-. ~KRÉM, ~PASZTA skosværte-
  n. ~SAROK skohæle-n,-e. ~TALP
  skosål-en,-e.
```

Angivelsen "(tsz)" opløses i forkortelseslisten til "többesszám" ('flertal'), mens "(tsz!)" henviser til deklinationsafsnittet i ordbogsgrammatikken.

2.3. Nielsen et al. (1982 og 1993)

Den skandinavisk↔ungarske leksikografi begynder for alvor at blomstre fra 1960'erne og fremefter.[3] Særligt er antallet af bilingvale ordbøger med ungarsk og svensk overraskende stort: Der findes os bekendt mindst 20 publicerede værker, af hvilke de vigtigste er György Lakó og Jozsef Fehérs *Svéd–magyar szótár* (Akadémiai Kiadó, 1969) på mere end 1.000 sider med 45.000 lemmata og 32.000 flerordsenheder, og Ferenc Kiefer og József Kiefers *Magyar–svéd szótár* (Akadémiai Kiadó, 1984) på 647 sider med 50.000 lemmata og 30.000 flerordsenheder. Førstnævnte udkom i fem udgaver (senest 1989), mens sidstnævnte kom i tre udgaver (senest 1992). På den baggrund, og når det tilmed tages i betragtning, at ungarsk som fag har kunnet studeres på Københavns Universitet fra 1967–1975 og på Aarhus Universitet siden 1972, er det lidt overraskende, at den første mellemstore bilingvale ordbog med dansk og ungarsk udkom så sent som 1982. Overraskende er det også, at der ikke findes noget spor af 1.-udgaven, dvs. af Nielsen et al. (1982), hverken i danske eller ungarske biblioteker. Hvor interessant det end havde været at vise et udsnit fra dette værk, som Nielsen et al. (1993:4f) betegner som "et pionerarbejde"[4], er vi desværre afskåret fra det og går direkte videre til andenudgaven, dvs. Nielsen et al. (1993).

Foruden øvrige præliminærsider indeholder Nielsen et al. (1993) følgende: "Forlæggerens forord"[5] på dansk (s. 4) og ungarsk

3 Den os bekendt tidligste bilingvale ordbog med ungarsk og svensk er János Lotz' *Ungersk–svensk ordlista* [...] fra 1940 (udg. af Ungerska Institutet vid Stockholms Högskola), mens den tidligste med ungarsk og norsk os bekendt er Dénes Csikys *Norvég–magyar magyar–norvég zsebszótár* [...] (Norsk Studentsamband i kommission hos Universitetsforlaget, 1957).

4 1982-udgavens eksistens er telefonisk blevet bekræftet af Margit Nielsen, men pga. hendes pludselige død har det ikke været muligt at få det fysiske værk i hænderne.

5 Det er faktisk uklart, hvorvidt der menes udgiver, dvs. ordbogsforfatterne, eller forlægger.

(s. 5); brugervejledning på dansk (s. 6) og ungarsk (s. 7); "Kort introduktion til det ungarske sprog" på dansk (s. 9–22); opslagsdel (s. 23–632); "Rövid bevezető a dán nyelvtanba" (dvs. 'Kort introduktion til den danske grammatik') på ungarsk (s. 633–638); litteraturliste (s. 639).

Termerne *monodirektional* og *bidirektional* bruges i faglitteraturen ofte i betydningen hhv. 'i hvilken der forefindes én L_a–L_b-lemmaliste' og 'i hvilken der forefindes såvel en L_a–L_b- som en L_b–L_a-lemmaliste'. I nærværende artikel skal der i overensstemmelse med Hausmann/Werner (1991) med førstnævnte term imidlertid forstås 'som retter sig mod brugerne af enten det ene eller det andet af værkets genstandssprog', mens der med sidstnævnte menes 'som retter sig såvel mod brugerne af det ene som brugerne af det andet genstandssprog'.

Intentionen om at være bidirektional antydes i Nielsen et al. (1993) dels med omslagets titelangivelse på begge sprog, dels med omteksten, som for størstedelens vedkommende ligeledes foreligger på begge sprog. Også forordet og brugervejledningen giver det indtryk, at ordbogen er konciperet bidirektional ("lige tilgængelig for både danske og ungarske brugere", s. 6f). Målgruppen er alle, som "vil kunne bruge [ordbogen] i deres arbejde, studier[6], etablering af nye kontakter og under gensidige besøg" (forordet, s. 4).

Brugervejledningen er meget kort; den indeholder en forkortelsesliste og oplyser bl.a. om, i hvilke former substantiver, verber og adjektiver figurerer i opslagsdelen. Den indeholder også en del fejl og sære formuleringer.

Introduktionen til det ungarske sprog er lang og forfejlet. Den består af en sær blanding af grammatik og sprogbrugsvejledning og er langtfra fejlfri (desuden gives oplysninger, som ret beset ikke hører hjemme i en ordbogsgrammatik: eksempelvis oplysninger vedr. sprogets oprindelse og udvikling). Der gives ganske vist en

6 Ordet "studier" (naturligvis på ungarsk) findes ikke i den ungarske version af forordet.

gennemgang af ordklasserne, men valget af grammatiske kategorier forekommer tilfældigt. For eksempel angives af de eksisterende 27 kasus i ungarsk først kun fem (tabel 2, s. 11), så seks (tabel 3, s. 12), hvorefter følger en liste over yderligere 12 kasussuffikser (tabel 4, s. 12) kaldet "De hyppigst forekommende endelser i præpositionsleddet" samt en opstilling af fem kasus i en sammenfatning (tabel 5, s. 13). Teksten bærer præg af, at forfatterne ikke er fortrolige med grammatikterminologi. F.eks. står "præpositionsled" (s. 12) anført som kasus, mens substantiv + possessivt suffiks kaldes "ejerfald"[7] (s. 20), og i stedet for den gængse grammatiske term *vokalharmoni* anvendes "medlydstilpasning" (s. 12), som i øvrigt også bruges i stedet for *assimilation*, og "lydharmonisering" (s. 18). Andre uortodokse og forvirrende betegnelser er "navnestedord" (s. 15) i stedet for *personligt pronomen* og "givefald" i stedet for *dativ*. Kapitlet om verber (s. 15–18) gengiver seks tabeller med konjugation, men angivelse af tempora og modi mangler helt i f.eks. tabel 8 (s. 15). På side 15 hedder det, at "hovedparten af de ungarske udsagnsord kan bøjes i både subjekt- og objektbøjning"; det gælder imidlertid kun de transitive verber, men forfatterne skelner slet ikke mellem transitive og intransitive verber.

Den korte indføring i dansk grammatik (s. 633–638) er mere gennemskuelig, men også her bruges idiosynkratiske termer: substantiv + artikel (f.eks. "pigen") betegnes som "bestemt fald" af substantiv, og artikel + substantiv ("en pige") som "ubestemt fald"; desuden hedder det, at "rene" adjektiver ("sød" og "sur") "angives i ordbogen i fælleskøn".

En betydelig mangel er det også, at der ikke findes en liste over de stærke og de uregelmæssige verber. En sådan liste ville have været hensigtsmæssig ikke kun for at undgå inkonsekvens, men også fordi den kunne hjælpe med verificering af ord i forbindelse med tekstlæsning.

7 Substantiv + possessivt suffiks er i ungarsk ikke genitiv (ejefald) men nominativ.

Trods blandingen mellem glatalfabetisk og nichealfabetisk opstilling er makrostrukturen forholdsvis enkel: Lemmata er opstillet glatalfabetisk, sublemmata nichealfabetisk, og komposita samt flerordsenheder anføres i sublemmaposition. I det følgende vises et udsnit fra opslagsdelen (s. 254f) gengivet så tro mod forlægget som muligt:

interims |-bevis -et, -er: részvényutalvány
►**løsning -en, -er:** közbenső megoldás
►**rapport -en, -er:** részjelentés
►**styre -t, -r:** ideiglenes vezetőség
[...]
interkontinental: (adj) földrészek közötti, interkontinentális
intermezzo -et, -er: közjáték
intern: (adj) belső
til ►t brug: belső használatra
[...]
interpolere {-ede}: (vb) közbeiktat, beszúr; [forfalske tekst] szöveget meghamisít; [matematik] interpoláció

En artikel er bygget op af følgende angivelsestyper: (1) lemma sat med fed type; (2) grammatisk angivelse (af enten bøjning (se **intermezzo**), ordklasse (se **interkontinental**) eller begge dele (se **interpolere**)); (3) betydningsdifferentiering (jf. inddelingen af ækvivalenter i **interpolere**-artiklen); (4) angivelse af en eller flere ækvivalenter. Nicheartiklerne, som indeholder sublemmata indledt med ► (jf. **interims |...**), er bygget op efter samme mønster.

Den valgte praksis er uhensigtsmæssig: Hvor er lemmaets begyndelse og slutning i f.eks. "**interims |-bevis -et, -er**"? Endvidere er placeringen af ordklasseangivelserne uheldig: Det er ikke entydigt (ved adjektiver og verber), om angivelsen gælder det pågældende lemma eller ækvivalenten. For substantivernes vedkommende er

der ingen tvivl, idet ordklassen her fremgår af bøjningsmønstret, som ligesom lemmata er sat med fed type (hvilket dog, som nævnt, giver indtryk af en forvirret lemmaforståelse).

De enkelte artiklers opbygning er mest konsekvent, når lemmaet er et substantiv. Ved verber angives de ungarske ækvivalenter i tredje person singularis præsens, men angivelsen af bøjningsmønstret er inkonsekvent: Ved de svage verber angives i de fleste tilfælde præteritum, men præteritum + perfektum participium forekommer også (f.eks. "**bebo** {**-ede, -et**}**:** (vb) lakik"). Ved de stærke verber angives præteritum + perfektum participium (f.eks. "**blive** {**blev, blevet**}**:** (vb) marad"), men præteritum alene findes også. Helt uigennemskuelige bliver angivelserne ved de uregelmæssige verber, hvor brugeren skiftevis præsenteres for hele bøjningsmønstret (f.eks. "**have** {**har, havde, haft**}", "**kunne** {**kan, kunne, kunnet**}") eller kun præteritum + perfektum participium (f.eks. "**være** {**var, været**}", "**skal** {**skulle, skullet**}"), hvor lemmaet undertiden angives i præsens i stedet for infinitiv, mens "**måtte** {**må, måtte, har måttet**}" har sit helt eget bøjningsmønster med ungarske ækvivalenter i præteritumsform.

De redaktionelle bemærkninger i form af betydningsdifferentieringen ved de nævnte verber er forkerte, idet **kunne**, **måtte**, **skulle** og **ville** er modalverber og ikke hjælpeverber. Desuden mangler der ofte en betydningsdifferentiering. F.eks. angives de følgende tre ækvivalenter for **have**: "bír, birtokol, van neki". Af disse kan det sidste altid anvendes, hvorimod "bír" udelukkende er litterært og arkaisk. Men ingen af ækvivalenterne kan bruges i eksemplet *Hvordan har du det?*, som oversættes med *Hogy vagy?*.

Omteksten fortæller intet om lemmaselektionen. Vi tør vove den påstand, at værket i forhold til sprogets mest almindelige gloser indeholder mange fremmedord og fagudtryk fra forskellige fagområder såsom administration, handel, landbrug og økonomi. Vi er desuden ikke overbeviste om, at værket virkelig kan bruges af både brugere med dansk som modersmål og brugere med

ungarsk som modersmål. Undertiden forekommer det endda, at ingen af disse to brugergrupper imødekommes. En bruger med dansk som modersmål vil f.eks. ofte have svært ved at vælge det adækvate udtryk, når der ikke findes redaktionelle kommentarer (foruden de allerede nævnte jf. f.eks. også "**grille -n, -r:** szeszély, hóbort, vesszőparipa, rigolya"). Samtidig får en bruger med ungarsk som modersmål heller ikke tilstrækkelig hjælp, når de redaktionelle kommentarer kun gives på dansk ("**gnist -en, -e:** [funke] szikra; [begejstring] lelkesedés; [levning] maradvány"). Men selv hvis man forstår kommentarerne, er de ikke nødvendigvis altid til hjælp, som det f.eks. er tilfældet med "[levning]" i netop nævnte eksempel, hvor der formentlig snarere menes 'en lille smule' som i udtrykket *gnist af håb*. Det rette ækvivalent er i så fald "szikrányi", som ikke figurerer i artiklen.

2.4. Nielsen et al. (1997)

Foruden øvrige præliminærsider indeholder bogen bl.a. et "Forlæggerens forord" på ungarsk (s. 6) og dansk (s. 7); en brugervejledning ligeledes på ungarsk (s. 8–9) og dansk (s. 10–11); "Kort introduktion til det ungarske sprog" på dansk (s. 12–25); liste over redaktionelle forkortelser (s. 26), opslagsdel (s. 27–838); "Rövid bevezető a dán nyelvtanba" (dvs. 'Kort introduktion til den danske grammatik') på ungarsk (s. 839–844); liste over de "mest hyppige" uregelmæssige danske verber (s. 845–847); samt liste over anvendt litteratur (s. 848).

Introduktionen til det ungarske sprog er stort set ordret overtaget fra Nielsen et al. (1993), den indeholder blot flere eksempler, en enkelt rettelse (hensynsfald) samt en ny tabel med titlen "Bøjning af navneord i ejefald samt ejefald + genstandsfald" (tabel 7, s. 17), som imidlertid næsten er uforståelig pga. forkerte grammatiske termer. Introduktionen til dansk grammatik er stort set også genbrug fra Nielsen et al. (1993), dog er afsnittet om verber

udvidet med oplysninger, som måske bedre havde hørt hjemme i brugervejledningen.

Ordbogens makrostruktur er forholdsvis simpel: Lemmata er opstillet glatalfabetisk, sublemmata nichealfabetisk, og flerordsenheder anføres i sublemmaposition. I det følgende vises et udsnit fra opslagsdelen:

adó [közteher] skat -ten,-ter; [rádió~] sender -en,-e; [adakozó] giver -en,-e; donor -en,-er; *adj* givende, givtig; ***~t kivet*** udskrive* skat; [...]
be|gyújt tænde+ op, fyre op; ►**gyullad** antændes+, gå* i brand; [pánikba esik] blive panikslagen; ►**gyűjt** indsamle; [tartozást] indkassere; ►**gyűjtés** indsamling -en,-er; indkassering -en,-er; ►**gyűjtő** indsamler -en,-e
[...]
fő *sb* hoved -et,-er; [személy] person -en,-er
fő *vb* koge+; ~ ***a feje*** han tænker hårdt

En artikel er bygget op af følgende angivelsestyper: (1) lemma sat med fed type; (2) angivelse af ordklasse sat med mindre typestørrelse (ved adjektiver og adverbier findes denne angivelse altid og ellers kun i tilfælde af homonyme ord tilhørende forskellige ordklasser, jf. **fő**); (3) evt. betydningsinddeling (jf. **adó**); (4) angivelse af en (eller flere) ækvivalent(er); (5) grammatiske data såsom angivelse af køn og pluralis af substantiver eller datid af svage verber efter følgende mønster: ingen markering for "-de, -ede", "+" for "-ete, -te" og "*" for uregelmæssige verber. Nicheartiklerne, som indeholder underartikler indledt med ► (jf. "►**gyullad**"[8] i artiklen **be|gyújt**), er bygget op efter samme mønster. Artiklen **be|gyújt**

8 Ækvivalenterne (f.eks. 'antændes') er forkerte, desuden er verbet ("begyullad") intransitivt, og betydningen 'gå betændelse i [ngt.]' mangler.

viser også, at det er ganske tilfældigt, hvad der får sublemmastatus (**gyújt** og **gyűjt** er ikke det samme).

Det er et fremskridt i forhold til de tidligere ordbøger, at angivelsen af verbernes bøjningsmønster er formaliseret og henviser til listen over uregelmæssige verber.

Ordforrådet i Nielsen et al. (1997) er udvidet dels inden for de fagområder, som vi har nævnt i forbindelse med Nielsen et al. (1993), dels med nye fagområder såsom informationsteknologi, medicin og sundhedsvæsen i det hele taget.

Intentionen om bidirektionalitet opretholdes med introduktionen til det ungarske sprog, men da grammatiske data kun gives ved de danske ækvivalenter, synes ordbogen primært at henvende sig til brugere med ungarsk som modersmål.

2.5. Det nye værk: Nielsen et al. (2008)

Nielsen et al. (2008) er en ny, revideret udgave af Nielsen et al. (1993), som er en anden og revideret udgave af Nielsen et al. (1982). Derfor er det vanskeligt at forstå, hvorfor udgiverne regner den nye ordbog for en 2.-udgave (jf. bogens titelblad og kolofon), da det vitterligt drejer sig om en 3.-udgave, såfremt der udkom en 1.-udgave i 1982. Allerede omslaget signalerer en ændring i forhold til 2.-udgaven: I stedet for de nationale farver har man valgt en neutral farvekombination, som den kendes fra de øvrige bilingvale ordbøger med eksotiske sprog fra Special-pædagogisk Forlag. Trods det lavere sideantal i forhold til 1993-udgaven er den nye bog mere omfangsrig; den har et større sideformat, mere tekst på siderne og en lidt mindre luftig opsætning.

Foruden titelblad og kolofon er der en fortekst (s. 3–6) og en opslagsdel (s. 7–533). Ordbogsgrammatikken fra 1993-udgaven – eller en forbedret version af den – mangler.

Forteksten indeholder en forkortelsesliste på dansk og ungarsk (s. 3), en tegnforklaring (s. 3) og en brugervejledning på dansk og

ungarsk (s. 4 og 5). Forkortelseslisten fortsætter overraskende nok på side 6 med en tabel dels over forkortelser, dels over interrogative pronominer. Selv om der ikke findes noget forord, som kunne oplyse om den påtænkte brugergruppe, kan vi gå ud fra, at også dette værk er tænkt bidirektionalt, bl.a. fordi forteksten er affattet på begge sprog.

Om brugervejledningen (både den danske og den ungarske) gælder først og fremmest, at den er misforstået: I et forholdsvis indviklet sprog og i en meget komprimeret form redegøres der for konjugation, deklination og komparation på ungarsk – noget, som ikke alene er umuligt på kun én enkelt side, men først og fremmest ikke hører hjemme i en brugervejledning, men derimod i en ordbogsgrammatik. Desuden er teksten mangelfuld. Eksempelvis hedder det også her ligesom i 2.-udgaven, at "mange verber har én bøjning med objekt og én uden objekt". Erstatningen af "subjekt- og objektbøjning" med det kortere udtryk "med/uden objekt" er uheldig, fordi den kan misforstås: Det er arten af objektet, der er afgørende for valget af bøjningen – den bøjning, som her kaldes "uden objekt", anvendes også i forbindelse med ubestemt objekt (f.eks. *Mit nézel?*). Det er heller ikke rigtigt, at "hjælpeverbet er/var" kun bruges i meget begrænset omfang på ungarsk (s. 4). For det første er der ikke tale om et hjælpeverbum i det anførte eksempel (*Den er rød*), men om et kopula, for det andet er det kun i 3. person singularis og pluralis præsens indikativ at prædikatet på ungarsk består af prædikativ + 0 finit verbum (jf. *piros* [= *den/det/han/hun er rød*]). På side 4 kan man læse, at passive verber i ungarsk "betragtes [...] som germanismer" og "bruges ikke i korrekt sprog". Faktum er, at ungarsk ikke har morfologisk markeret diatese.

Afsnittet om substantiver er så overfladisk, at det lige så vel kunne have været udeladt. Der findes tre eksempler på deklination i en tabel, som angiver fem kasus (betegnelsen "præpositionsled" for kasus er i øvrigt bibeholdt). Det nævnes ikke, at tabellen kun repræ-

senterer en lille del af kasussystemet. En anden mangel er bindevokalerne: Da de ungarske substantiver gennemgående angives i nominativ singularis, ville det have været hensigtsmæssigt at vise hele rækken af mulige bindevokaler foran pluralis "-k" (dvs. "-a-", "-e-", "-o-", "-ö-") i stedet for kun de to, som fremgår af eksemplerne.

Også den ukommenterede tabel over pronominer på dansk og ungarsk (s. 4) er problematisk: Det, som står under overskriften "Refleksivt stedord", er ikke refleksive pronominer men akkusativformen af personlige pronominer.

Vejledningen for ungarske brugere (s. 5) behandler ligeledes verber, substantiver og adjektiver. Her finder brugeren mere præcis information om lemmaformerne samt om, hvilke grammatiske oplysninger ordbogen giver om de enkelte ordklasser: For verbernes vedkommende er det infinitivformen, der figurerer i lemmaposition, hedder det, desuden angives datid af svage verber[9] med "*{-ede}*" eller "*{-te}*" og efterfølgende ordklassen (for verbers vedkommende "*vb*"). Ved stærke verber angives både datid og perfektum participium (f.eks. "**skrive** *{skrev, skrevet} vb*").

Det er beklageligt, at forfatterne kalder de stærke verber for uregelmæssige verber og derved helt glemmer de uregelmæssige verber, hvilket bl.a. medfører, at de optræder to gange i opslagsdelen og i øvrigt med inkonsekvente angivelser; jf. f.eks. "**kunne** *{kunnet, præs: kan} vb*", "**kan** *{kunne, kunnet} vb*", "**må** *{måtte} vb*" og "**måtte** *{har måttet; præs: må} vb*". En liste over de uregelmæssige verber mangler stadig.

Ved substantiver angives køn, pluralis og ordklasse ("**brev -et, -e** *sb*"). Men også her findes diskrepans mellem tekst og tabel, idet flertalsendelsen på dansk ifølge forfatterne er "***-r***". I tabellen finder man imidlertid eksempler på de forskellige flertalsendelser ("***-e***", "***-er***", "***-r***", "-"). Hvorvidt flertalsendelserne "*-e*", "*-er*" og "-" betragtes som uregelmæssige, forbliver uklart for brugeren.

9 Verber, der ender på hhv. *-ede* og *-te*, kaldes hhv. 1. klasse og 2. klasse, men inddelingen i de to klasser har ingen konsekvens i ordbogen.

Adjektiverne angives med bøjningsendelserne og ordklassebetegnelse ("*adj*"). Desuden fremgår det af tabellen (s. 5), at både præsens participiums- og perfektum participiumsformer har lemmastatus, men intet om principperne.

Det er en betydelig mangel, at man intetsteds erfarer noget om ordbogens brugerafgrænsning, ligesom man heller intet får at vide om dens funktioner. Efter vores vurdering imødeses følgende kommunikative brugersituationer:

- oversættelse (fra dansk til ungarsk)
- reception (af tekster på dansk)

Hvorvidt der med 'oversættelse fra dansk til ungarsk' skal forstås oversættelse fra L_1 til L_2 eller oversættelse fra L_2 til L_1, er uklart. Da værket ikke indeholder nogen ordbogsgrammatik med ungarsk som genstandssprog og der i øvrigt kun gives få grammatiske data i artiklerne, kan værket ikke hjælpe brugere med dansk som modersmål i forbindelse med tekstproduktion på ungarsk, medmindre det drejer sig om brugere, som i forvejen behersker ungarsk. Funktionen 'oversættelse fra dansk til ungarsk' forstået som hjælp til oversættelse fra L_1 til L_2 er derfor ikke understøttet optimalt.

Direktionaliteten er uklar og diskutabel: Efter vores vurdering kan værket bedst bruges af brugere med ungarsk som modersmål. Således imødekommes brugssituationen 'oversættelse dansk→ungarsk' forstået som oversættelse $L_2 \rightarrow L_1$ ofte på tilstrækkelig vis; jf. eksempelvis:

format *-et, -er*, *sb* [størrelse/form] alak és méret; formátum; [personlig] képesség, rátermettség, ütőképesség; [it] formátum; ***lommeformat*** zsebformátum

Efter (1) lemmaet følger (2) bøjningsmønster ("***-et, -er***"), (3) angivelse af ordklasse (i dette tilfælde "*sb*") og (4) ungarske ækvi-

valenter inddelt efter betydning. Ved artiklens udgang ses et (5) sublemma uden angivelse af grammatiske data, men med en ækvivalent. Det (6) polyseme leksems betydninger inddeles i "[størrelse/form]", "[personlig]" og "[it]", hvilket er hensigtsmæssigt i forhold til funktionen 'oversættelse dansk→ungarsk' og i forhold til brugere med ungarsk som modersmål. Det er uvist, hvorvidt funktionen 'tekstproduktion på dansk' konceptuelt er forudset. Understøttet er den i hvert fald ikke.

Uheldigt og inkonsekvent i forhold til den mindre kyndige bruger med ungarsk som modersmål er det, at de redaktionelle kommentarer i artiklerne ikke gennemgående gives på ungarsk eller, såfremt ordbogen virkelig skal kunne tjene såvel danske som ungarske brugere, på begge sprog. I artiklen **format** gives kommentarerne kun på dansk (i kantet parentes), mens de i artiklen **ligge** gives på såvel dansk som ungarsk (førstnævnte i kantet, sidstnævnte i almindelig parentes):

> **ligge** *{lå, ligget} vb* fekszik; [høne] kotlik, költ (tyúk); [militær] állomásozik (katonaság); ***ligge forrest*** élen van, vezet (versenyben); ***ligge godt i hånden*** jól símul a kézbe (pl. szerszám); [...] ***det ligger lang tid tilbage*** ez már régen volt; [...]; ***fejlen ligger i at*** a hiba abban rejlik, hogy; a hiba ott van, hogy; [...]; ***min beslutning ligger fast*** döntésem megmásíthatatlan; ***priserne ligger fast*** az árak nem változnak; [...]

Efter (1) lemmaet følger bøjningsformer i krøllet parentes og (2) ordklasseangivelse. Herefter følger (3) ungarske ækvivalenter og (4) en lang række flerordsenheder på dansk med (5) angivelse af ækvivalente udtryk på ungarsk. Efter ækvivalentangivelsen til det almindelige "ligge" (*fekszik*) ses (6) betydningsdifferentieringen "[høne]" og "[militær]". At differentieringen stedvis angives på begge sprog – "[høne]"/"(tyúk)" hhv. "[militær]"/"(katonaság)" –

kunne tyde på en tilstræbt bidirektionalitet; i resten af artiklen er de redaktionelle kommentarer dog kun på ungarsk.

Det er problematisk, at flerordsenheder på dansk ikke findes i deres grundform, men i stedet i selvlavede eksempelsætninger. Havde man – for bare at nævne ét tilfælde – i stedet for eksempelsætningen ***priserne ligger fast*** anført kollokationen *ligge fast*, havde det været lingvistisk mere korrekt og sprogpædagogisk mere hensigtsmæssigt. Som artiklen ser ud nu, er det ikke kollokationer, idiomer osv., der oversættes, men eksempelsætninger, som undertiden end ikke tilnærmelsesvis ækvivalerer med de respektive flerordsenheder på målsproget.

Makrostrukturen udgøres af en glatalfabetisk lemmaliste med initialalfabetisk præsentationsform og er således mere brugervenlig end 1993-udgaven, som indeholder en blanding af glatalfabetisk og nichealfabetisk ordning. Enkelte steder findes rester fra 2.-udgavens nichealfabetisering, f.eks. sublemmaet ***lommeformat*** i artiklen **format**.

Hensigtsmæssigt i forhold til tilgangsstruktur er det, at ikke kun lemmata, men også samtlige danske flerordsenheder er sat med fed type og i kursiv. Uhensigtsmæssigt i forhold til tilgangsstruktur er det til gengæld, at der tilsyneladende ingen henvisninger findes.

3. Konklusion

De eksisterende bilingvale ordbøger med dansk/ungarsk falder ind under kategorien folkeleksikografi. Det kan derfor diskuteres, hvorvidt det giver mening med fuldt akademisk beredskab at påvise alskens mangler, uhensigtsmæssigheder og konceptuelle problemer. På den anden side kan der også sættes spørgsmålstegn ved det meningsfyldte i at afstå fra en videnskabeligt funderet kritik.

Meningen med de tidlige ordbøger/ordlister – dvs. Világi (1945), DBK (1957) og Knudsen (1958) – har aldrig været, at de skulle bruges til decideret tekstoversættelse, tekstproduktion og/eller tekstreception. De er produkter af bestemte historisk-politiske omstændigheder og skulle tjene som værktøjer til at afhjælpe *basale* kommunikationssituationer i hverdagen. Det er en banal og inadækvat konstatering, at værkerne med et lemmaantal svingende fra 2.250 til 3.700 per lemmaliste er for små til at kunne yde tilstrækkelig hjælp til tekstoversættelse, tekstreception og tekstproduktion. Mere adækvat er det at kritisere Nielsen et al.-værkerne, som i forhold til deres ambitionsniveau fremviser en fattig struktur samt talrige konceptuelle problemer såsom uklar direktionalitet, uklar brugerafgrænsning og uklar funktionalitet.

Bibliografi

Leksikografiske opslagsværker

DBK 1957: *Dansk–ungarnsk (lille ordbog)/Magyar–dán (kis szótár)*. København: Danske Boghandleres Kommissionsanstalt.

Knudsen, Jørgen 1958: *Dán–magyar és magyar–dán szótár. Dansk–ungarsk og ungarsk–dansk ordbog*. Kbh.: Mellemfolkeligt Samvirke.

Nielsen, Margit/Ágnes Sitkeiné Szira/József Szira 1982: *Dansk–ungarsk ordbog. Dán–magyar szótar*. Első kiadás. Alsóőrs: Északi Fény BT.

Nielsen, Margit/Ágnes Sitkeiné Szira/József Szira 1993: *Dansk–ungarsk ordbog. Dán–magyar szótar*. Második, bővített kiadás [2. reviderede udgave]. Alsóőrs: Északi Fény BT. [Genoptrykt i 1998].

Nielsen, Margit/Ágnes Sitkeiné Szira/József Szira 1997: *Magyar–dán szótár. Ungarsk–dansk ordbog*. Első kiadás. Alóőrs: Északi Fény BT.

Nielsen, Margit/Ágnes Sitkeiné Szira/József Szira 2008: *Dansk–ungarsk ordbog. Dán–magyar szótar*. 2. udvidede udgave, 1. oplag. Herning: Special-pædagogisk Forlag.

Világi, Zoltán 1945: *Dán–magyar magyar–dán kis szótár*. Vostrup: Majholm (Jylland).

Anden litteratur

Hausmann, F.J./R.O. Werner 1991: Spezifische Bauteile [...]. I: F.J. Hausmann, O. Reichmann, H.E. Wiegand, L. Zgusta (Hrsgg.): *Wörterbücher: Ein Internationales Handbuch zur Lexikographie* [...]. Berlin/New York: Walter de Gruyter, III, 2729–2769.

Sørensen, Søren Peder 2005: *De ungarske soldater. En glemt tragedie fra den tyske besættelse af Danmark under 2. verdenskrig*. Museet for Varde By og Omegn.

Tarp, Sven 2004: Hvad er en bilingval ordbog? I: *LexicoNordica* 11, 5–22.

Tarp, Sven 2006: *Leksikografien i grænselandet mellem viden og ikke-viden. Generel leksikografisk teori med særligt henblik på lørnerleksikografi. Bind I–II*. Århus: Center for Leksikografi.

Loránd-Levente Pálfi
forskningsassistent
Center for Leksikografi
Handelshøjskolen,
Aarhus Universitet
Fuglesangs Allé 4
DK-8210 Aarhus V
llp@asb.dk

Erzsébet Stokholm
Lillegrund 85
DK-2300 København S
stokholmer@mail.dk

Sven Tarp
professor
Center for Leksikografi
Handelshøjskolen,
Aarhus Universitet
Fuglesangs Allé 4
DK-8210 Aarhus V
st@asb.dk

En SAOB-artikel växer fram

Bo-A. Wendt

This article presents a specific example of how the work of an editor of the comprehensive historical Swedish dictionary, published by the Swedish Academy, (SAOB) proceeds. The example concerns the Swedish preposition *under* 'under' and accurately describes the process of writing that SAOB entry. Drawing from this particular word, it is shown how etymology governs also the semantic description, on what principles the different meanings of a word is either divided into different main sections or kept together within one and the same main section, how SAOB deals with borderline cases, how the description interacts with already written parts of the dictionary, how the main sections may be grouped together in bigger subgroups, how subordinate sections are distinguished and furthermore how the use of a word in compounds may influence on the description of the simplex word.

Det är många personer inblandade i det arbete som ligger bakom en artikel i *Ordbok över svenska språket* utgiven av Svenska Akademien, vanligen kort och gott SAOB – den stora historiska ordbok över det nysvenska (skrift)språket som började ges ut redan 1893 och vars senast tryckta band (bd 35) når fram till och med verbet *tyna*. Den första arbetsinsatsen ligger i regel mycket långt före den egentliga redigeringen och utgörs av excerpering ur det digra källmaterialet. Avsevärt närmare artikelns slutförande ligger sedan en kompletterande excerpering (inte minst ur andra ordböcker) och en förberedande redigering med huvudsakligt sikte på morfologiska och ortografiska variabler (se Vifot 1975, Tjebbes 1993 och om ett specifikt exempel Persson 2002). Därnäst följer formredovisningens inarbetande i en fullständig artikel, varvid också uttal,

etymologi och framförallt betydelser och användningar, åtföljda av dokumenterande och exemplifierande språkprov, skall beskrivas. Ännu har emellertid artikeln långtifrån nått sin fullbordan, utan sedan följer punktuella utlåtanden från sakkunniga inom olika ifrågakommande fackområden, kontroll av citat och hänvisningar (se Starfelt 1975) samt ett antal genomläsningar och granskningar av innehåll och formalia (av andra än den ursprunglige författaren, mest avgörande av försteredaktör och ordbokschef) och så allra sist en tryckteknisk bearbetning.

I det följande ämnar jag litet närmare beskriva endast en del av denna process, nämligen de vägval som redaktören gör från det att vederbörande tar över ansvaret för ett ord (närmast från den förberedningsassistent som kartlagt dess morfologi och stavning) och till dess att vederbörande lämnar ifrån sig en första version av artikeln (närmast till innehållsläsning och kontroll). Det handlar alltså bara om ett visst skede i en SAOB-artikels framväxt, inte om hela dess tillblivelse alltifrån de anlag som ligger nedlagda hos de från början enbart kronologiskt ordnade excerptlapparna ända till den slutliga förlösningen i ett nytryckt häfte. Förvisso handlar det om ett ytterst avgörande skede i fosterutvecklingen, om jag tillåts hålla kvar den frammanade bilden, men alls inte om alla de krafter som är med om att prägla slutprodukten. Det handlar inte ens om alla insatser härvidlag som den enskilda artikelns redaktör förväntas stå för, endast om dem vederbörande gör i den första vändan, ensam på sin kammare (eller snarare: sitt tjänsterum), innan kollegernas och chefens iakttagelser, synpunkter och invändningar kommer med i spelet.

Till detta kommer att den följande beskrivningen är helt och hållet uppbyggd kring ett specifikt exempel. Den avser alltså en enda SAOB-artikel, inte mer allmänt hur redigeringsarbetet brukar gå till. Det ord det därvid kommer att handla om tillhör till yttermera visso inte språkets i antal olika ord räknat stora ordklasser utan tvärtom en av de mest slutna formordsklasserna. Jag

ämnar nämligen beskriva framväxten av artikeln **UNDER**, preposition och adverb, – som tillhör, eller rättare sagt: kommer att tillhöra, en jämförelsevis mycket exklusiv typ av artiklar i SAOB (jfr med Wendt 2002:182). Likafullt vill jag bestämt tro att både valet att beskriva bara ett exempelord – med alla dess specifika enskildheter – och valet av exempel ger möjlighet att säga inte så litet av mer eller mindre generell räckvidd för hur SAOB-artiklar brukar växa fram (dock givetvis med den reservationen att olika redaktörer går till verket på litet olika sätt). Det blir alltså ett slags begränsat studiebesök under någon månad i SAOB-verkstaden hos en och samme redaktör, av vilket man ändå förhoppningsvis kan lära sig ett och annat om verksamheten där överhuvudtaget.

Beskrivningen tar fasta på redaktörens bakomliggande resonemang och överväganden, varför motspänstigheter i materialet får det huvudsakliga utrymmet, medan det som utan större besvär låter sig fogas in i artikelmallen – och sådant finns det trots allt rätt mycket av också – kommer i skymundan i det följande. Den fortsatta framställningen är indelad i två huvudavsnitt: ”Etymologi och former” och ”Betydelsebeskrivning”, av vilka det senare är det ojämförligt mest omfattande. Detta motsvarar väl omfångsförhållandena i en normal SAOB-artikel, där betydelsebeskrivningen vanligtvis får det allra största utrymmet. Detta avsnitt 2 sönderfaller i ett antal delavsnitt vari jag visserligen tematiskt behandlar olika sidor av betydelsegrupperingen var för sig men ändå försöker kronologiskt följa framväxten av artikeln steg för steg.

1. Etymologi och former

Etymologi är på många sätt utgångspunkten för SAOB:s beskrivning och avgörande redan för vad som skall bli egna artiklar. Ett formords flerfaldiga ordklasstillhörighet och användning som till exempel både preposition med rektion och adverb föranleder

dock aldrig uppdelning i olika artiklar. Vad gäller **UNDER** ger de etymologiska handböckerna däremot entydigt vid handen att vi egentligen har att göra med två olika indoeuropeiska ord, vilket normalt ofrånkomligen ger två artiklar i SAOB. För flertalet betydelser, däribland alla nu levande hos det självständiga ordet, är visserligen ursprunget ett och detsamma, en komparativbildning till en rot med betydelsen 'nere, nedanför' eller något liknande. Därutöver finns det emellertid en användning av *under* som har annat ursprung. Den lever nu bara som förled i några sammansättningar och är lånad från tyska i de nordiska språken. Mot ett *under* med en grovt angiven grundbetydelse av 'nedanför' står ett lån med betydelsen 'bland; mellan', som återfinns i bland andra *underhålla* och *underrätta*. Den hör ytterst samman med till exempel latinets *inter* (möjligen av ett annat avljudsstadium). Samma rot finns också i det dialektala substantivet *undar*(*n*) 'mellanmål'. I (väst)germanskan har dessa båda olika arvord mycket tidigt fallit samman till sin form.

Frågan är då om dessa olika ursprung bör återspeglas i SAOB:s orduppdelning, så att vi skulle få två artiklar: det inhemska 'nedanför' med mera i **UNDER**, prep.[1] och det lånade 'bland; mellan' i **UNDER**, prep.[2]. I en beskrivning av tyska kunde man ju tycka att det skulle ha fog för sig, men det är inte gjort så i DWB (bd 11:3, sp. 1472 ff. (moment II B)), inte heller i den danska ODS (bd 25, sp. 1166 f. (moment 15)). Också för SAOB:s räkning har jag funnit för gott att hålla samman betydelserna under ett **UNDER**; när man en gång lånade in uttryck och sammansättningar där *under* betydde 'bland; mellan' i svenskan, gjorde man det utan tvekan som ett betydelselån, inte ett ordlån. (Hur detta sedan skall fogas in i artikelns momentindelning skall jag återkomma till senare.)

I övrigt ter sig etymologin, liksom för den delen den form- eller stavningsvariation som förekommer, rätt enkel och rättfram; redaktören har i stort sett bara att skriva av, eller rättare sagt: stöpa om i SAOB-form, det som skäligen samstämmigt sägs i de

gängse etymologiska handböckerna. Det enda egentliga besväret med etymologiparentesens utformning är att det som nästan alltid har sina vanskligheter att fast- och säkerställa de indo-iranska frändernas transkriberade form med stöd av primärlexikon som ibland bara tillhandahåller dem otranskriberade. När detta väl är avklarat och förhållandet mellan de två etymologiskt obesläktade betydelserna är utrett, är så artikelns etymologiparentes klar. Av formvariationen är det väl främst varianter utan slut-*r* som är litet intressanta, men med dem är det inte mer att göra än att redovisa de kronologiska spann med belägg som vi har.

Riktigt så här enkelt kommer emellertid SAOB-redaktören inte undan etymologin, ty denna är sedan också styrande för betydelsernas inbördes ordning i artikeln. Och här uppstår i **UNDER**:s fall genast nästa delikata – etymologiskt grundade – spörsmål. Som jag sade, är *under*, i sin vanliga användning, ytterst en komparativbildning och alltså i grund och botten äldst adverbiellt. Skall då dess adverbiella användningar i SAOB:s material placeras före de prepositionella i momentordningen? Så är gjort i den tyska DWB (bd 11:3, sp. 1455 ff.), men eftersom de adverbiella *under* åtminstone i svenskan rätt entydigt är sekundära till de prepositionella, förefaller en sådan momentföljd missvisande här. Vårt adverbiella *under* är knappast en arvtagare till det gamla indoeuropeiska utan yngre än det prepositionella. För fornvästnordiskt skaldespråk hävdas det visserligen i LexPoet (s. 580) att *und* är prepositionellt alltmedan *undir* ”bruges postpositivt og som adv” (på samma sätt som *eft* : *eftir* med flera, enligt en antagen allmän skillnad mellan dessa en- och tvåstaviga former; jfr med Johannisson 1943:60); samtidigt heter det emellertid att ”i hds bruges i reglen *undir*, men metrum viser oftest det rette”, och de adverbiella belägg som anföres (i ett avslutande moment C, s. 581) har huvudsakligen tycke av absolut använda prepositioner. Ulf Teleman (2008:77) hänför utan vidare *under* till ”det äldsta skiktet av germanska prepositioner” och antar som adverbbaserade källförbindelser till andra yng-

re prepositioner istället en sammanställning av adverb + (äldre) preposition (s. 111 ff.; om *under*:s äldre adverbiella bakgrund se s. 112 f.). Jag har – för övrigt i gott sällskap med Söderwall (bd 2: 814 ff.) – stannat för att användning som preposition med utsatt rektion får bli moment **I**, användning som absolut preposition eller adverb moment **II**.

2. Betydelsebeskrivning

Med etymologin avklarad vidtar det lappsorterande som är redaktörens lott vid betydelseanalysen. Den kronologiskt ordnade mängden av excerptlappar med ett ord skall nu sorteras i högar av betydelser och underbetydelser. I **UNDER**:s fall rör det sig om sammanlagt 2 106 lappar på huvudordet. Det gör det till ett rätt stort ord men inte till de verkligt stora i SAOB-redigeringen. Verbet **SÄTTA**, substantivet **TID** och prepositionen **TILL** omfattade till exempel alla uppemot eller gott och väl 10 000 excerpter vardera. Vilka de entydigaste betydelserna bör bli har man vanligtvis rätt klart för sig redan innan sorteringen börjar på allvar – med stöd i andra ordböckers indelning. Därvid behöver man inte ha bestämt sig för exakt vilka huvudbetydelser man vill urskilja; också undermomentens lappar skall ju till sist hamna i var sina högar, så det är förstås mycket vunnet om man redan från början har sorterat dessa för sig. Sedan får man givetvis kallt räkna med omflyttningar mellan högarna när den precisare betydelsebeskrivningen väl tar vid. Då gäller det nämligen att noga se till att den definition man omsider formulerat täcker allt man har i sin momenthög – och därutöver inget som hamnat annorstädes bland högarna. Det kan också hända att någon i och för sig distinkt uttryckstyp som man i förstone lagt i en egen hög visar sig vara belagd med så få lappar att ett eget undermoment inte synes av nöden, varvid dessa lappar helt enkelt får sorteras in i den större hög som företräder

närmast överordnade moment. Att man å andra sidan ur högar som blivit stora efter den första grovsorteringen så småningom plockar ut mindre högar med sidoordnade undermoment är, av naturliga skäl, väl så vanligt. Att den första sorteringen står sig ända in i artikelns färdiga momentindelning händer förstås bara vid ord med högst begränsat antal lappar – och dit hör förvisso inte ett ord som *under*.

Den allra största delen av SAOB:s material på **UNDER**, prep. o. adv., faller inom det som enligt ovan utgör moment **I**, vanlig prepositionell användning med en utsatt rektion, mer än fyra femtedelar av lapparna. Det är också här som de avgörande besluten med avseende på momentindelning får göras, eftersom moment **II** sedan, i första hand åtminstone – enligt ovan antagna sekundära ställning –, bara är tänkt att återspegla samma betydelsestruktur i annan syntaktisk funktion.

2.1. Samma eller olika huvudmoment?

Det är skäligen uppenbart att den grundläggande betydelsen är den som grovt kan beskrivas som 'nedanför'. Man kan emellertid fråga sig huruvida skillnaden mellan denna betydelse och den litet snävare 'nedanför och inunder' (med ursäkt för cirkeldefinitionen härvidlag) utgör två så åtskilda betydelser att de borde redovisas var för sig, rentav som två huvudmoment i SAOB:s struktur. Efter en del funderande fram och tillbaka bestämde jag mig för att de inte gör det utan att de gott kan slås samman. Jag tror nämligen att skillnaden huvudsakligen är avhängig av rektionsreferentens väsen. Skillnaden mellan *under fönstret* och *under taket* har nog ytterst mest att göra med att tak har en större vågrät undersida under vilken någon eller något naturligt kan befinna sig i sin helhet, medan fönster på en vägg inte har det. Några fall där en betydelseskillnad hos själva *under* kan läsas in går det nog ändå att hitta: *under huset* avser vanligtvis en placering inne under husets grund,

men det kan ju också handla om en befintlighet längre ned i en sluttande terräng. En sådan åtskillnad till trots nöjer jag mig likväl här med ett tillägg i huvudmomentet under **1**: ”äv. med inbegrepp av läge där det lägre belägna också befinner sig förskjutet i sidled i förhållande till det högre belägna”.

När det gäller att skilja ut vad som är egna huvudmoment försöker jag leva efter en tumregel som rentav blivit något av en käpphäst för mig. När nära varandra liggande betydelse(nyanse)r föreligger, prövar jag gärna två huvudbetydelsekandidater mot varandra genom att se huruvida det är möjligt att få fram båda betydelserna i alldeles samma satssammanhang. Är det det, har vi, menar jag, starka skäl att antaga två så distinkt olika betydelser att de bör redovisas i var sitt huvudmoment. Det man prövar är ju nämligen om betydelseskillnaden är oberoende av sitt textsammanhang – om den bärs upp av ordet självt och inte bara är en följd av att det används i olikartade sammanhang. När en betydelseskillnad föreligger i sådana helt likalydande satser, har vi alltså att göra med ett slags (lexematiska eller sememetiska) minimala par.

Jag uppmärksammades på testets bevisvärde när jag skrev på en annan prepositionsartikel (**TILL**) för flera år sedan men har sedan med framgång tillämpat det också på innehållsord. Den exempelmening med prepositionen *till* som var klargörande för mig och som faktiskt sedermera kommit med bland redaktionsexemplen på ömse håll var: *Du kan använda de där lådorna till blomkrukor* med två olika betydelser beroende på vad *till* betyder. Alltså har vi här rimligtvis två olika huvudbetydelser, om än båda finala: funktion (vad något är avsett som, ofta alltså ’såsom’) kontra ändamål (vad något är avsett för, ofta alltså ’för’). I detta fallet kan man förvisso hävda att betydelseskillnaden också styrkes av de olika parafraserna, *såsom* kontra *för*. Ofta framhålls också sådana parafraseringar som avgörande i urskiljandet av olika betydelser, ibland med indragande också av parafraser på andra språk. Rent

principiellt menar jag emellertid att ”minimala-par”-testet är det bättre och säkrare av de båda vid urskiljande av huvudbetydelser, eftersom det parafraserande ordet kan ha en annorlunda strukturerad betydelse, som kan medföra just delvisa överlappningar. I all synnerhet får man se noga upp med parafraser på andra språk, så att man inte lägger detta främmande språks egen betydelsestrukturering av den utomspråkliga verkligheten som ett normerande mönster, ett facit, över svenskans. För att engelskan skiljer på *ceiling* och *roof* har givetvis inte svenskans *tak* två olika betydelser motsvarande de engelska ordens, för att ta ett övertydligt exempel.

För att återvända till vårt *under* har jag under grundbetydelsen **1** också valt att redovisa både sådana bildliga användningar där *under* tillämpas på ett värde på en skala eller måttstock och liknande och betyder ’lägre l. mindre l. ringare än’ (till exempel *alla under fem år*, *under sin förmåga*) och sådana utvidgade användningar där tanken på ett lägre läge i lodled är undanskymd (till exempel *under bilden finns en bildtext* eller *under ekvatorn*). Här ger själva textsammanhanget upphov till respektive särbetydelse, samtidigt som också den rent bildliga användningen ännu har så tydligt kvardröjande koppling till den rent rumsliga att den bör beskrivas såsom underordnad denna, inte som en självständig huvudbetydelse.

Tydligt egen är däremot betydelsen ’innanför’; detta styrks inte bara av den (inomspråkliga) parafrasen *innanför* motsatt *nedanför* i moment **1** utan också av ”minimala-par”-testet: *under jackan skymtade en grön tröja* kan antingen handla om en påklädd person eller om en hög av kläder på golvet. Jag vill emellertid också som egen betydelse urskilja en som i viss mån påminner om (och nog har vuxit fram ur) den nyss dryftade *under huset*-användningen. Det handlar om en betydelse ’nedanför och invid eller vid foten eller randen av’. Den är klart ovanligare än de helt centrala ’nedanför’ och ’innanför’ och används nu nästan bara i sådana uttryck som *under berget* eller *under land* (till sjöss). Att den bör skiljas ut

som en egen huvudbetydelse visas inte bara av den annorlunda parafrasen utan också av ”minimala par”: *de bodde under berget* kan betyda antingen ’inne under berget’ (i till exempel en grotta) eller ’vid foten av berget’. Ja, det går till och med att skapa sig en mening där alla de tre nyssnämnda huvudbetydelserna kan bli aktuella: *under muralmålningen fanns en hällristning* – där hällristningen kan finnas antingen en bit nedanför på samma lodräta yta (’nedanför’) eller på en vågrät häll framför denna yta (’vid foten av’) eller (helt fräckt, får man ju säga) övermålad på densamma (’innanför’).

2.2. Gränsfallsformler

Den mellersta av dessa betydelser får bli mitt huvudmoment **2**. Den står trots allt närmare ’nedanför’-betydelsen än vad ’innanför’ gör, ja, uppvisar till och med ett antal äldre språkprov i materialet där det nog ännu finns kvar ett visst mått av en föreställning om att det omtalade förutom att vara framför eller invid rektionsreferenten också befinner sig något lägre än denna. Det här verkar ju misstänkt likt det fall jag tidigare dryftade, *under huset*-typen ovan. Skillnaden skulle alltså vara att det i det förra fallet är betydelsen ’nedanför’ som är den centrala, medan det i det senare fallet istället trots allt är betydelsen ’framför’ som är viktigast. Jag har anfört denna användning som ”utan klar avgränsning från **1**”. Ett sådant, idag mindre brukligt, exempel är detta:

> Han umgås endast med Spinhuus-Dockor, sätter möte med dem under min näsa. (1740)

I den mån vi idag alls skulle använda exempel av det här slaget skulle vi läsa det som ’nedanför’ snarare än ’framför’, i anslutning till den ännu levande *under huset*-typen under moment **1**.

Också mellan det som enligt ovan förts till moment **1** (’ne-

danför') och moment **3** ('innanför') finns det emellertid litet besvärliga gränsfall. Det gäller sådana prepositionsfraser som *under mattan* och *under täcket*, som av SO (s. 3369) beskrivs som särfall under betydelsen 'innanför' "med tonvikt på döljande o. d.". Förvisso är döljandet ofta en mycket framträdande komponent i dessa fall, inte minst i bildliga användningar av uttrycken (som *sopa ngt under mattan*, *spela under täcket*). Jag vill likväl mena att tanken på ett huvudsakligt lodledsförhållande är tillräckligt stark för att hålla kvar frastypen under **1**, särskilt som det finns belägg där ett döljande bara är en alldeles ofrånkomlig följd av rektionsreferentens väsen. (Jämför också med *under jackan*-exemplet ovan, som tydligt kan fördelas på **1** och **3**, men där användning också i betydelse **1** naturnödvändigt innebär något slags döljande.) Det blir litet konstigt om ordets användning om samma rumsliga verklighet skulle räknas olika betydelser till godo beroende på hur väsentlig en bakomliggande avsikt av döljande är. Tack och lov finns det i SAOB:s formeluppsättning också bot för sådana här gränsfall. Jag har helt enkelt garderat mig under **1** med ett "stundom närmande sig eller övergående i **3**", särskilt tillämpligt på bildliga uttryck som *sopa ngt under mattan*, *spela under täcket* o. d. (*med ngn*) – som jag placerat i ett undermoment omfattande rektioner som betecknar "(skyddande eller täckande eller döljande) hölje o. d.". Man får trösta sig med att sådana här gordiska hugg är ofrånkomliga när den alltid lika brokiga språkbruksverklighet som fått sitt avtryck i lappmaterialet skall sorteras in i fasta momentkategorier.

2.3. I ordboken redan behandlade uttryck

Alla likalydande prepositionsfraser måste förstås inte nödvändigtvis föras till varsitt ena och samma moment, inte ens när det handlar om snarlika betydelser. Välbelagda i materialet är användningar av *under* med *hand* och *namn* som (huvudord i) sin rektion, och dessa fördelar sig på flera olika betydelsemoment. Fra-

serna är redan behandlade i SAOB på respektive substantiv, vilket både underlättar och kan försvåra. Det underlättar givetvis väldeliga därigenom att jag nu på **UNDER** bara behöver avgöra var i dess betydelsestruktur som en hänvisning till respektive ställe skall stå och sedan luta mig mot redan förefintlig betydelsebeskrivning och beläggsredovisning. Det kan därmed emellertid också ställa till det en del, eftersom behandlingen i substantivartiklarna förstås i första hand tagit fasta på vad substantivet betyder och därigenom kan ha kommit att sambehandla fraser som till äventyrs uppvisar litet olika betydelse hos *under*.

I artikeln **NAMN** upptas användningar med *under* i ett par olika undermoment under dess första grundläggande huvudmoment. I den mån dessa uttryck tar sikte på ett vilseledande bruk av ett namn, vill jag uppfatta *under* i dem som hörande till ovan beskrivna moment **3**. Här blir det sålunda en hänvisning till **NAMN**, dock med tillägget ”stundom utan klar avgränsning från **5** *b* γ eller ε”. Till dessa åberopade undermoment under **5** (= underordnings-*under*, varom mera nedan) hör sedan användningar där man genom namns uttryckliga nämnande tillskriver någon något som i *ge ut en bok under sitt eget namn* (en delmängd av ett av undermomenten under **NAMN**) respektive där något innehållsligt ingår som en del av eller sammanfattas av en större eller (begreppsligt) etiketterande enhet som i *en förening under namn av Verdandi* (en delmängd av det andra undermomentet under **NAMN**).

För uttrycket *under hand* är det omvänt så att samma betydelse av *under* är förhanden i flera av de förekomster som redovisas på olika håll i artikeln **HAND**, nämligen sådana där *under* snarast har sin grundläggande betydelse **1** (’nedanför’). Därför blir det därunder ett litet undermoment med bara hänvisningar. När man som här bara hänvisar till en annan artikel som står för hela beskrivningen av uttrycket, brukar jag annars välja att göra det ihop med annan definitionstext. Eftersom en sådan hänvisning fråntar (ja, rentav berövar) en kravet på beläggande språkprov, inkräktar det

nämligen inte på det utrymme man har till sitt förfogande i ett moment för att belägga dess innehåll. Att göra en sådan hänvisning till ett eget undermoment är alltså onödigt – eller rättare sagt bara nödvändigt om man tycker sig vilja kunna hänvisa entydigt till det från något annat ställe i artikeln. Och det är just det som är skälet till att denna endast hänvisande redovisning av användning med rektionen *hand* fått bli ett eget språkprovstomt undermoment. Lösningen underlättar nämligen korshänvisning hit från de här åberopade parallellställena i artikeln **UNDER** där användning med *hand* som rektion tas upp (samma uttryck, annan betydelse).

2.4. Momentgruppering ovanför huvudmomenten

Kommen så här långt i min momentuppdelning insåg jag att jag nu satt med tre huvudmoment som skulle kunna sammanfattas som rumsliga, vilket jag därför satte in som ett **A** ovanför momenten **1–3**. Detta fick emellertid följder för den fortsatta momentordningen. Dittills hade jag tänkt mig att de mer abstrakta betydelserna av *under* skulle få följa och så allra sist den etymologiskt självständiga betydelsen 'bland eller mellan'. När nu detta rumsliga **A** kommit till stånd, ställdes detta på ändan. Också sistnämnda betydelse är ju rumslig eller åtminstone därav nära avhängig. Jag kunde därför inte gärna göra annat än att sätta in denna (nu stendöda) etymologiskt främmande betydelse sist under **A**, istället för allra sist. Momentet, som därmed blir nummer **4**, inleds med en etymologisk parentes av lydelsen "efter motsv. anv. av mlt. *under* l. t. *unter*". Alldeles stendöd är nog inte betydelsen, ty det torde vara spår av denna vi hittar i sammansättningarna *därunder*, *härunder* och *varunder* i uttryck av typen *X, varunder Y och Z återfinns*. Detta klaras av med en inledande bruklighet "†, utom i *b*" och så hänvisningar till de sammansatta adverben **DÄR-**, **HÄR-** o. **VAR-UNDER** under detta *b*.

Därefter får då ett **B** följa, omfattande bara ett enda huvudmo-

ment: moment **5** som omfattar det vi kan kalla underordnings-*under*. Efter detta **B 5** följer så ett moment **I** avslutande bokstavsmoment **C** där det handlar om parallellitet och liknande mellan ett skeende eller förhållande och ett annat vilket senare anges i rektionen. Här var jag från början inne på att allt det som SO (s. 3369 f.) redovisar i sammanlagt fyra olika moment skulle kunna bakas ihop till ett enda – utifrån tanken att skillnaden i huvudsak beror på typen av rektion, inte på *under* i sig. I SO föreligger varsitt huvudmoment för typen *under medeltiden* (moment 3 där), typen (*komma*) *under behandling* (moment 5), typen (*arbeta*) *under tystnad* (moment 6) och typen *under ngn förevändning* (moment 7). Det samfällda skulle då kunnat beskrivas som att ett skeende eller förhållande kännetecknas av (samtidighet med) det som anges av rektionen. Detta skulle då kunna omfatta också den tidfästande användningen. Omsider beslöt jag mig likväl för att särhålla denna i ett eget huvudmoment. Den speciella funktionen att tidfästa något görs ju med förankring i en tidsram som ligger utanför det tidfästa skeendet (rektionen därför i typfallet i bestämd form, angivande något redan känt), medan de andra nyssnämnda *under*-användningarna alla beskriver något som är en del av själva skeendet (rektionen därför i typfallet i obestämd form, angivande något som introduceras med den övriga skeendebeskrivningen).

Dessutom, eller snarare: just därför, kunde jag också hitta stöd i mitt tidigare ”minimala-par”-test. Vanligtvis består ju rektionerna i tidfästningsfallen och de andra fallen väsensenligt av helt olika slags uttryckstyper, men i det konjunktionella *under det att* kan vi få fram båda betydelserna; i *hon arbetade under det att krampanfallen ansatte henne* kan bisatsen vara antingen tidsadverbiell eller sättsadverbiell, med litet olika innebörd. I det senare fallet hör dess *under* då till det huvudmoment som fått bli **6** (och det första under **C**), som omfattar fall där en handling eller ett skeende eller tillstånd (till sitt väsen) kännetecknas av det som anges med rektionen. I det förra fallet hör dess *under* istället till det som fått bli mo-

ment **7**, omfattande fall av samtidighet med det i tiden avgränsade skede som anges med rektionen: vanligtvis tidfästande som i *under medeltiden*, men i vissa fall angivande varaktighet som i *under tre timmar*. Dessa två betydelsevarianter är rätt snyggt fördelad på bestämd och obestämd form hos rektionen: *under veckan* : *under en vecka*, varför det näppeligen finns skäl att överväga ytterligare uppdelning på olika huvudmoment.

2.5. Undermomentsindelning

Indelningen i undermoment är nästan alltid ett mycket friare val än indelningen i huvudmoment. För dessa senare är den semantiska verkligheten helt avgörande (eller bör åtminstone vara det!), men när man väl, på förhoppningsvis goda och obestridliga grunder, urskilt sig ett huvudmoment, är valet av urskilda undermoment genast mycket mer beroende av vad man mest tycker sig vilja fokusera i sin beskrivning – och därmed inte sällan en ren smaksak, där olika lösningar kan vara lika rimliga och förenliga med ordets semantik. Det man får vara beredd på är att varje sådant val av överordnad indelningsgrund nästan alltid också har sina nackdelar och ibland leder till onaturliga splittringar längre ned i undermomentstrukturen. Ett tydligt exempel på detta är mitt val att tudela nyssnämnda moment **5** i användning med personbetecknande rektion (*a*) och med sakbetecknande rektion (*b*). Valet grundar sig på att dessa båda rektionstyper överlag, var sitt väsen trogen, ger upphov till litet olika varianter av underordning. Med personbetecknande rektion handlar det i första hand om ställning i en beslutshierarki (som i *att tjänstgöra under ngn*), med sakbetecknande gärna om mindre reglerade beroendeförhållanden (som i *slav under begäret*).

Så välgrundad tudelningen är, har den sina baksidor. I ett underundermoment *a* α redovisas lydnad och dylikt gentemot en överordnad person, medan den lydnadsplikt som följer av underord-

ning under en organisation eller liknande – genom sin rektionstyp – hamnar under *b* (i ett *b* α). Till detta senare undermoment har jag också, med strikt tillämpning av den överordnade indelningsgrunden men med reservationen ”äv. med anslutning till *a* α”, fört rektioner som (metonymiskt) betecknar eller utgör symboler för överhet(sperson), täckande uttryck av typen *någon annan under svensk krona hörande stad* (1765). Om jag istället valt att urskilja lydnads- eller tjänandeförhållande på den översta undermomentsnivån, hade jag i gengäld fått sprida ut person- och sakrektionerna på de olika undermomenten, åter och åter görande reda för när bara endera eller båda är tillämpliga. Ett alternativ vore förstås att föra den för animat överhet metonymiska sakbeteckningen till mitt *a* α med en formulering i stil med ”äv. i användning som motsvarar *b*”, men jag tyckte att det vore att förfuska den övergripande indelningen något. Dessutom skall det i ärlighetens namn sägas att en följdrätt renodling i person- respektive sakrektioner gör det hela litet lättare för redaktören. Jag slipper ju därmed sitta och grubbla över när de formellt sakliga överhetsbeteckningarna i så fall vore tillräckligt smyganimata för att rymmas i ett utvidgat *a* α: *svenska kronan* – ja, förmodligen, *regimen* – kanhända, *finansdepartementet* – nja ...

Också i moment **6** har jag infört en uttömmande tudelning på översta undermomentsnivån, här – litet fyrkantigare – efter syntaktisk bundenhet: syntaktiskt nödvändiga i *a* (någorlunda svarande mot SO:s moment 5: *komma under behandling*) och syntaktiskt fria i *b* (någorlunda svarande mot SO:s moment 6+7: *vara sövd under behandlingen*, *gå under förevändningen att* ...). Under **6** *a* faller inte bara sådant som *vara under behandling*, *ha läget under kontroll* eller *hålla under uppsikt* utan också uttryck som *under vapen* och (med kvardröjande rumslig betydelse) *under segel*, alltså fall där *under* styr en konkret rektion som mer eller mindre metonymiskt får stå för ett skeende eller förhållande. Denna variant återkommer under *b*, bland de syntaktiskt fria bestämningarna

(som i *göra ngt under tårar*), så här står jag åter med ett fall där en indelningsgrund skär tvärs igenom en annan.

2.6. Syntaktiska moment med romerska siffror

Med momentet **7**, tids-*under*, är så hela moment **I** genomskrivet. I jämförelse med nusvenska definitionsordböcker brukar SAOB för det mesta komma upp i flera moment än de på motsvarande ord där – naturligt nog när också döda eller halvdöda betydelser finns med i leken och beskrivningen är så mycket mer detaljerad. I detta fallet har den preliminära SAOB-artikeln jämnt lika många huvudmoment under **I** som prepositionen *under* har i SO. I SAOB:s beskrivning ryms då också två (varav en död) som saknas i den senares. I gengäld har jag, som jag nyss visat, valt att samredovisa tre av SO:s huvudmoment i ett enda. Övriga fyra moment är mer eller mindre helt parallella.[1] Artikeln **UNDER** är nu inte klar med detta; förutom den synnerligen vidlyftiga pendangen med sammansättningar, som jag alltså inte kommer att säga något alls om i detta sammanhang, återstår ännu moment **II** om absolut eller adverbiell användning av *under*. Som redan sagts är emellertid den grundläggande semantiska analysen här klar på förhand, eftersom de adverbiella användningarna förväntas motsvara de prepositionella. Detta faller också ut rätt snyggt och prydligt; de prepositionella betydelser som saknas bland de adverbiella är, inte helt

1 Det finns i dagens SAOB-skrivande nog överlag en tendens att tillskapa färre huvudmoment än vad som en gång varit brukligt, också oberoende av det tilltagande kravet på kortare artiklar överhuvudtaget. Ordboken har därför av allt att döma blivit mer hierarkisk och mindre linjär i sin betydelsebeskrivning (jfr Malmgren 1992: 138). En föregångsman härvidlag torde ha varit förre ordboksredaktören Carl-Erik Lundbladh, som var min handledare när jag började på ordboken och den som eftertryckligast har präglat min syn på ordboksskrivandet – inräknat denna strävan efter hierarkisk artikelstruktur. I sin SAOB-artikel **SÅ**, adv., konj., pron., interj., urskiljer Lundbladh endast fem adverbiella huvudmoment, i **TAGA**, v., endast sju transitiva huvudmoment – rätt påtagligt sparsmakat i jämförelse med många artiklar med liknande mångtydiga ord i äldre band.

oväntat, de i moment **I 2**, 'vid foten av', och **I 7**, tids-*under*. Inga huvudbetydelser utan motsvarighet under **I** dyker upp, så moment **II** stannar alltså på fem huvudmoment.

Utöver genomgången av de rätt få excerptlapparna som hör till moment **II** innebär arbetet nu också en genomgång av det omfattande beståndet av redan skrivna särskilda förbindelser med ett verb + *under*. Eftersom långtifrån allt material därifrån kommit med också i mitt, blir det till att noggrant se till att dessa förbindelser inte uppvisar äldre förstabelägg (eller yngre sistabelägg vid döda betydelser) än de jag redan har i kapslarna. En hel del kompletteringar av detta kronologiska slag blir det också tal om, innan moment **II** är fullständigt. I övrigt kommer emellertid detta moment i mångt och mycket att utgöras av hänvisningar till respektive särskild förbindelse, varför jag i min artikel kan utelämna de tillhörande språkproven. Vissa distinkta betydelsegrupper under den överordnade huvudbetydelsen kan på det viset bilda helt språkprovstomma undermoment – så länge detta inte innebär att jag därigenom döljer något som skulle varit det äldsta på hela huvudbetydelsen. Sålunda urskils under **II 1** (”motsv. **I 1**”) ett *a* som avser införflyttande eller tillskapande av något under något annat som stöd eller underlag och så följer ett tiotal verb med hänvisningar till var sina särskilda förbindelser och därefter inga språkprov alls.

Ett problem som återkommer emellanåt vid hänvisningarna till särskilda förbindelser är att de där anförda språkproven inte alltid uppvisar *under* i helt entydig partikelfunktion (utan kanhända istället i användning som preposition med rektion). Överlag har jag stillatigande varit lojal med ordbokens tidigare analyser, men i något fall ändå ansett en brasklapp vara på sin plats: ”utan klar avgränsning från [motsvarande prepositionella moment under **I**]”. Det gäller inte minst om de verbförbindelser med reflexivt pronomen såsom objekt som redovisas under **II 4**, ”motsv. **I 5**” och alltså avseende underordnings-*under*, med hänvisningar till

en lång rad som särskilda förbindelser beskrivna uttryck, alltifrån **BRINGA UNDER SIG** till **TAGA UNDER SIG**.

Genom att mycket av beläggsredovisningen således är hänskjuten till andra artiklar i ordboken, och genom att den precisare betydelsebeskrivningen är hänskjuten till moment **I**, kan hela moment **II** klaras av på rätt litet utrymme.

2.7. Användning i sammansättningar

När man så sitter med en första version av sin momentindelning, väntar lapparna med det som inom redaktionen kallas andraledssammansättningar, de sammansättningar vari ens ord utgör senare led och som alltså i ordboken finns kringspridda i alfabetsräckan. För vissa ord, i synnerhet substantiv av mer allmänt slag, kan dessa ibland vara uppemot ett hundratal, för en preposition som *under* är antalet däremot högst beskedligt. (Här är det istället de nyssnämnda särskilda förbindelserna med *under* som fått god spridning i tidigare band av ordboken.) De andraledssammansättningar som finns beskrivna i ordboken, teoretiskt sett också övriga, förväntas man nu placera ut på sin olika betydelsemoment under huvudordet, uttryckligen redovisat för ofta bara ett urval av dem – efter ett *jfr* efter respektive moments språkprov. Mestadels blir detta ett sätt att bara ytterligare belägga ordets användning i förhandenvarande betydelse, men genomgången av dessa andraledssammansättningar blir ju också ett prov på hur väl man i sin betydelsebeskrivning fått med alla relevanta betydelse(skiftninga)r och beskrivit dem på ett fullödigt sätt. Särskilt när sammansättningarna är många händer det inte så sällan att genomgången föranleder vissa ändringar eller, kanhända oftare, tillägg i den föreliggande beskrivningen.

Ett generellt problem vid denna betydelsesortering är att sammansättningarnas ordboksdefinitioner till sin kärna, naturligt nog, innehåller just det ord man själv sitter med – utan närmare

precisering av vad det avses betyda just här. Redaktören som en gång i tiden skrev definitionen kunde ju inte gärna förutse vilken betydelsestruktur som i tidens fullbordan skulle tillkomma detta och har inte heller så ofta haft uppenbara skäl att problematisera dess flertydighet. När då definitionen inte åtföljs av några utförda språkprov, bara nakna (det vill säga med endast källhänvisning), blir det ibland svårt att av ordbokstexten allena avgöra exakt vilken av efterledens omsider beskrivna betydelser som avsetts.

Vad gäller **UNDER** har jag för **DÄR-**, **HÄR-** och det ännu oskrivna **VAR-UNDER** helt enkelt valt en placering omedelbart efter momentet **I**:s övergripande text, före alla siffermoment. Det är gjort efter mönster av andra prepositionsartiklar och utgör ju en rimlig generalisering, eftersom de teoretiskt bör kunna hänföra sig till vilket som helst av de prepositionella momenten (också om inte allt råkar vara belagt i sammansättningsmaterialet). Av övriga sex andraledssammansättningar, alla med adverbiella förleder, fördelar sig fyra på minst två olika huvudbetydelser hos **UNDER**. På **IN-UNDER** och **NED-UNDER** finns en momentindelning som delvis går på tvärs med min indelning. På **IN-UNDER** sambehandlas sålunda förflyttning in under något (del av mitt **I 1**) med det att något kommer i någons ägo eller under någons välde (del av mitt **I 5**), medan allt detta hålls skilt från befintlighet under något (också i mitt **I 1**) – i sin tur däremot sammanhållet med befintlighet innanför något (mitt **I 3**). Förvisso används ordet *under* i dessa definitioner, men så pass entydigt och dessutom så pass väl kompletterat med utförda språkprov att det inte innebär några större problem att så att säga översätta deras betydelsemoment till mina egna för det osammansatta *under*.

Tillräckligt starka skäl för att i någon mån ändra på dessa senare till bättre överensstämmelse med de förras litet annorlunda strukturering har jag å andra sidan inte tyckt mig finna. Jag har nöjt mig med att portionera ut dem på rätt moment enligt min egen betydelsestruktur, varvid flera som framgått kommit att åter-

finnas under flera olika huvudmoment. På **IN-UNDER** föreligger alltså en momentåtskillnad mellan förflyttning och befintlighet, något jag själv egentligen aldrig övervägde på allvar för det enkla *under*. Visst kan man tycka att ordet har litet olika betydelse i *flytta* (*in*) *under taket* och *sitta under taket*, men i enlighet med min allmänna grundhållning vill jag främst, kanhända helt och hållet, tillskriva de olika satssammanhangen denna skillnad. (Faktiskt är nog skillnaden då större mellan *under taket* ('nedanför o. inunder') och *under fönstret* (bara 'nedanför'), som jag ju enligt ovan också avstod från att särhålla.) Möjligen är en momentuppdelning som på **IN-UNDER** mer motiverad i kasusspråk som formellt särhåller rörelse och befintlighet, såsom hos fornsvenskans *undir*. Söderwall (bd 2: 814 ff.) skiljer mycket riktigt på översta momentnivån mellan ett **I** "med dat." och ett **II** "med ack." (dessutom ett **III** "med gen."), och sedan återfinns den grundläggande betydelsen 'nedanför' som varsitt moment **1** på ömse håll under de två första romerska momenten. Också i DWB (bd 11:3, sp. 1457 ff.) görs dativ- gentemot ackusativrektion till överordnad indelningsgrund. Å andra sidan hör ju skillnaden i kasusval hemma hos rektionen; det är dess form som reglerar prepositionens betydelse, och några "minimala par" enligt min modell kan det därför aldrig bli tal om.

Också det beskrivna ordets användning som förstaled i sammansättningar[2] (liksom för den delen i avledningar eller för ett verb i särskild förbindelse med partikel) utgör givetvis ett viktigt korrektiv för betydelsebeskrivningen av huvudordet. Jag kommer som redan framhållits inte att säga något vidare om dessa förstaledssammansättningar med *under*, otaliga i jämförelse med de få andraledssammansättningarna; i skrivande stund är de i själva verket ännu obeskrivna. Eftersom en hel del av dem är verb som har sin omedelbara motsvarighet bland de särskilda förbindelserna (*skriva under* – *underskriva*), redan beaktade till och med verb på T, känns ändå det återstående mötet med alla de många *under-*

2 Om urvalet av sammansättningar i SAOB se Stille (2003).

sammansättningarna inte som något överhängande hot mot den hittills frammejslade momentindelningen.

2.8. Sakkunnigfrågor

Ett korrektiv av alldeles annat slag utgör svaren på de frågor man efter genomskriven artikel sänder iväg till redaktionens därför särskilt kontrakterade sakkunniga inom olika fackområden. De kan ibland, särskilt i substantivartiklar, innebära en hel del korrigeringar av framförallt bruklighetsuppgifter och definitionernas utformning. Hos ett formord som *under* är däremot inslaget av fackspråk högst begränsat. Det enda som föranledde mig att rådfråga sakkunniga gällde användningar till sjöss, i uttrycken *under* {*land*/*ett fartyg*/*ett sjömärke*} och liknande samt *Under!* såsom varningsrop när man kastar ned något från riggen. Eller rättare sagt: kastade, ty sakkunningsvaren bekräftade det som materialet pekade på, att detta varningsrop inte längre är i bruk. Också för den förstnämnda, ännu gängse, uttryckstypen stämde för övrigt sakkunnigbedömningen med det jag tyckt mig kunna utläsa ur vårt material.

3. Slutord

Den här redovisade beskrivningen av prepositionen *under* gäller för en SAOB-artikel som ännu är kommen bara en bit på vägen. Inte endast vad de många sammansättningarna med *under* kan komma att innebära för beskrivningen är ställt under framtiden (för att ta till en numera rätt marginell *under*-användning!), det gäller också de ändringar, stora eller små, som artikeln kan komma att undergå när den – efter färdigställandet av en första fullständig version – kommer under ögonen på de kolleger som har att innehållsgranska, citat- och hänvisningskontrollera och så till sist

korrekturläsa den. Dessa arbetsmoment föranleder inte så sällan en hel del omstöpningar och förändringar, numera oftare nedstrykningar än tillägg. När jag likväl lämnat dem utanför den här redovisningen av en SAOB-artikels framväxt, beror det på att jag velat fokusera på de vägval och överväganden redaktören gör i sin första och grundligaste brottning med materialet. Trots all knådning och omarbetning som ibland kan återstå innan en artikel når sin slutgilla form, är det ändå detta första skede som med få undantag satt den väsentligaste prägeln på det som omsider finner väg till trycket. Det är också det skede då redaktörens arbete är som mest ensamarbete, vilket, inbillar jag mig, torde göra det till det kanhända minst genomskinliga, ja, rentav mest okända, i hela arbetsprocessen vid SAOB:s redaktion.

Läsaren har här bara fått ett enda exempel från en enda redaktörs arbete. För den intresserade återstår att utröna många andra aspekter av och inslag i denna del av det redaktionella arbetet med det svenska språkets fylligaste lexikala kartläggning. Likväl avslutar vi med detta studiebesöket i SAOB-verkstaden för denna gången, och jag drar mig åter ensam tillbaka på mitt arbetsrum därsammastädes.

Litteratur

Ordböcker

DWB = *Deutsches Wörterbuch*, von Jacob Grimm und Wilhelm Grimm, hrsg. von der Deutschen Akademie der Wissenschaften zu Berlin, Band 1–16 + Quellenverzeichnis, Leipzig 1852–1971.

LexPoet = *Lexicon poeticum antiquæ linguæ septentrionalis* / *Ordbog over det norsk-islandske skjaldesprog*, opr. forfattet af Sveinbjörn Egilsson, 2. udg. ved Finnur Jónsson, København 1931.

ODS = *Ordbog over det danske Sprog*, grundlagt af Verner Dahlerup, udg. af Det Danske Sprog- og Litteraturselskab, bind 1–28 + supplement (bind 1–5), København 1919–1956, 1992–2005.

SO = *Svensk ordbok*, utg. av Svenska Akademien, utarbetad vid redaktionen för Svenska Akademiens samtidsordböcker, Lexikaliska institutet, Institutionen för svenska språket, Göteborgs universitet, 2 band, Stockholm 2009.

Söderwall, K. F. 1884–1918: *Ordbok öfver svenska medeltids-språket*, band 1–2 + supplement, utg. (1925–73) av K. F. Söderwall, W. Åkerlund, K. G. Ljunggren och E. Wessén, Lund.

Övrig litteratur

Johannisson, Ture 1943: Afton. I: *Meijerbergs arkiv för svensk ordforskning* 5, utg. av Styrelsen för Meijerbergs institut vid Göteborgs högskola, 50–75.

Malmgren, Sven-Göran 1992: Om behandlingen av polysema ord i enspråkiga ordböcker. I: *Nordiske studier i leksikografi. Rapport fra Konferanse om leksikografi i Norden 28.-31. mai 1991*, red. av Ruth Vatvedt Fjeld, Skrifter utg. av Nordisk forening for leksikografi 1, Oslo, 137–143.

Persson, Christina 2002: Den förberedande redigeringen av ordboksartikeln TID. I: *Alla ord är lika roliga. Festskrift till Lars Svensson 28 februari 2002*, Stockholm: Norstedts, 127–141.

Starfelt, Gulli 1975: Något om citatkontrollen vid Svenska Akademiens ordboksredaktion. I: *Kring en ordbok. Festskrift till Sven Ekbo 7 augusti 1975*, Stockholm: Norstedts, 122–128.

Stille, Per 2003: Sammansättningar i ordboksarbetet, med utgångspunkt från TEATER. I: *Nordiske studier i leksikografi 6. Rapport fra Konference om leksikografi i Norden, Tórshavn 21.–25. august 2001*, red.: Zakaris Svabo Hansen & Anfinnur Johansen, Skrifter udg. af Nordisk forening for leksikografi 7, Tórshavn, 341–349.

Teleman, Ulf 2008: Tidiga nordiska prepositioner. Härkomst och uppkomst. I: *Arkiv för nordisk filologi* 123, 77–141.

Tjebbes, Gunilla 1993: Den förberedande redigeringen av SAOB. I: *Ord och lexikon. Festskrift till Hans Jonsson 10 juni 1993*, Stockholm: Norstedts, 195–199.

Vifot, Eva 1975: Hur ordboksartiklar förbereds. I: *Kring en ordbok. Festskrift till Sven Ekbo 7 augusti 1975*, Stockholm: Norstedts, 195–202.

Wendt, Bo-A. 2002: Prepositionernas grammatik i SAOB. I: *Alla ord är lika roliga. Festskrift till Lars Svensson 28 februari 2002*, Stockholm: Norstedts, 181–192.

Bo-A. Wendt
fil. dr, ordboksredaktör
Svenska Akademiens ordboksredaktion
Dalbyvägen 3
SE-224 60 Lund
Bo.Wendt@nordlund.lu.se

ANMELDELSER

Ordbok över karelskan på Internet

Ilse Cantell

1. Ordboken

Webbversionen av *Karjalan kielen sanakirja* (förk. *KKS*, sv. Ordbok över karelskan) lades ut på Internet den 15.10.2009. Personerna bakom webbordboken är Marja Torikka som redaktör[1] och Jari Vihtari som planerare av webbgränssnittet. Ordboken ligger på adressen http://kaino.kotus.fi/cgi-bin/kks/kks_etusivu.cgi.

Den tryckta ordboken består av sex band och gavs ut 1968–2005. Satu Tanners presentation *En dokumentation av det karelska språket* finns i LexicoNordica 13 (2006:169–183). Här tar jag därför bara fram sådant som är specifikt just för webbordboken.

I en intervju som är gjord den 13.5.2005, då den tryckta ordboken kom ut, uttrycker Marja Torikka sin önskan om att ordboken ska digitaliseras. ”Det blir lättare att forska utgående från den. Om man t.ex. har behov av att få tillgång till dialekten i en viss kommun, kan den inte lätt extraheras ur boken, men maskinellt skulle det gå lätt och bekvämt.” (Miikkulainen 2005). Nu har Torikkas önskan gått i uppfyllelse.

Webbversionen av KKS (nedan kallad *Webb-KKS*) är en del av en omfattande samling texter i gratistjänsten *Kaino* på Internet (http://kaino.kotus.fi/) som Forskningscentralen för de inhemska språken tillhandahåller för alla som t.ex. i forskningssyfte är intresserade av det finska språket och dess släktspråk, såsom karelskan och samiskan.

1 Marja Torikka var redaktionssekreterare för den tryckta KKS fr.o.m. 1988.

2. Processen från handsatta boksidor till webbordbok

Överföringen av den karelska ordboken till elektroniskt format har skett något olika beroende på hur gammalt ordboksmaterial det är fråga om.

De tre första banden som utgavs 1968–1983, handsattes i tiden på traditionellt sätt. Materialet i dessa tre band, dvs. bokstavsintervallet A–N har först skannats från de tryckta sidorna. De skannade filerna har sedan gått igenom en textigenkänningsprocess (OCR), dvs. överföring från bild till text med typografisk formatering.

De tre sista banden, bokstavsintervallet O–Ö kom ut under åren 1993–2005 och för deras del var manuskripten skapade och sparade i elektronisk form som formaterade WordPerfect-filer. Efter textigenkänningen förelåg de äldre och de nyare manuskripten i samma form, och resten av processen var identisk för dem alla. Den fortsatta processen bestod i en konvertering till XML-kodad och strukturerad data och sedan en rent teknisk konvertering så att ordbokens lexikografiska kategorier presenteras som de ska i det för ordboken planerade användargränssnittet. Användargränssnittet och de olika sökfunktionerna presenteras i avsnitt 4 nedan.

Innehållet i de sex banden är så gott som oförändrat med undantag av några korrigerade felaktigheter och, för det första bandets del, ett visst förenhetligande som var nödvändigt för att artiklarna ska motsvara redigeringsprinciperna i de senare banden.

3. Hur *Webb-KKS* ser ut och fungerar

3.1. Tekniska krav på utrustningen

Att olika operativsystem fungerar olika är naturligtvis en utma-

ning för planeringen av en webbordbok och det är troligen omöjligt att gardera sig mot all variation i fråga om system och inställningar som användarna kan ha.

Materialen innehåller en mängd specialtecken och för Kaino är alla tecken kodade enligt standarden Unicode. De många specialtecknen i Webb-KKS återges enligt användarinstruktionerna bäst om användaren har gratisfonten Charis SIL installerad. Fonten finns för både Windows, MacOSX och Linux.

3.2. Användargränssnittet

3.2.1. Megastruktur och design

Webb-KKS är till sin megastruktur en hel webbplats, vars delar dels innehåller material från omtexterna i den tryckta ordboken, dels sådant material som är specifikt för webbordboken.

Förutom den egentliga ordboksdelen ingår i *Webb-KKS* en presentation av den tryckta ordboken och dess relation till webbordboken, instruktioner för sökningar och ett pdf-dokument som innehåller den tryckta ordbokens inledning på ca 100 sidor. Författare till inledningen är ordbokens första huvudredaktör Pertti Virtaranta. I den tryckta ordbokens inledning ingår en hel del foton som inte ännu finns med i inledningen till *Webb-KKS*. Det är synd men troligen kan de skannas in inom en snar framtid. I fotona ingår mycket kulturhistoria som via publicering kunde få spridning via Internet.

Bland materialet på webbplatsen finns en karta över det språkområde ordboken täcker och en förteckning av dialekterna (som förkortningar och utskrivna) grupperade och markerade med olika färger[2]. De nordkarelska dialekterna är markerade med mörkblått, de sydkarelska med rött och de olonetsiska dialekterna med grönt. Dessa färger återkommer i de geografiska markörerna i ordartiklarna.

2 Kartan kan ses på adressen http://kaino.kotus.fi/kks/kkskartta.html.

Som ett interaktivt element finns i webbordboken en länk till ett formulär för respons från användarna. Arbetet med att utveckla sökmöjligheterna i webbordboken pågår fortfarande på Forskningscentralen för de inhemska språken. Avsikten är att bl.a. göra det lättare att skriva in dialektala sökord med specialtecken eller diakritiska tecken.

Webb-KKS har en grafisk design som anknyter till den tryckta ordboken genom att varje webbsida (med undantag av den karelska dialektkartan) till vänster i ordboksfönstret har ett fält som ser ut som ryggen på den tryckta ordboken.

I användarinstruktionerna har man varit mycket noggrann och sett till att användaren får all möjlig hjälp. Man varnar också för vissa brister som ännu inte åtgärdats. Bl.a. påpekar man att (exempel)sökningarna kanske inte alltid ger ett fullständigt dataunderlag. I sådana fall instrueras användarna att klicka sig fram till de enskilda artiklarna där de kan komplettera resultatet. En förbättring av detta är under arbete och motiverad.

3.2.2. Ordboksfönstret

Ordboksfönstrets högra två tredjedelar upptas i själva ordboken av ett sökformulär med var sin flik för de två olika sökmöjligheterna artikelsökning och exempelsökning. Under formuläret ses sökresultaten.

Ett röd-svart-guldfärgat fält återkommer på webbplatsens andra sidor och länkar till webbplatsens andra delar är placerade i fältet. Samma länkar finns också överst i ordboksfönstret. Eftersom ordboken är en vetenskaplig ordbok och i mitt fall en ordbok över ett språk jag inte behärskar är behovet av användarinstruktioner åtminstone i början rätt stort. Dessutom går det att konsultera en snabbhjälp som öppnas som ett extrafönster och finns bakom en länk på sökformuläret.

En artikel i *Webb-KKS* är indelad i avsnitt med radbyten. Detta strukturerar artikeln tydligare än vad fallet är med den tryckta ord-

boken där det inte har varit möjligt på grund av det naturliga kravet att spara utrymme. Uppslagsordet med ordklassmarkör finns på en egen rad, och därefter står betydelsenumren på egen rad om de åtföljs av ett exempel eller en dialektvariant av uppslagsordet.

De dialektala varianterna av uppslagsordet och exempelfraserna finns sedan i ett eget, ofta ganska stort block som det också gärna kunde ha varit indelat i kortare avsnitt. Men någonstans måste ju gränsen dras och risken finns att man i stället för strukturering får splittring.

Exemplen är i stället separerade med lodstreck och efter ett exempel finns en geografisk markör för socknen där exemplet är upptecknat. Markörerna bär färgkoder och är i de flesta fall förkortade. Då man för kursorn till en sockennamnsförkortning, öppnas med hjälp av mouseover-funktionen ett mycket litet extrafönster med det utskrivna sockennamnet. Att också de sockennamnen som inte alls är förkortade är försedda med ett extrafönster med exakt samma text är onödigt, men hellre så än inga extrafönster alls. I stället kunde också andra förkortade markörer ha försetts med utskriven form i ett extrafönster.

3.3. Sökningar

3.3.1. Sökning i ordlistan och sökning med hjälp av sökformulär

Då man går in i ordboken ser man i ordboksfönstrets vänstra del ett röd-svart-guldfärgat fält vars synligaste del är en lista med 21 uppslagsord i alfabetisk ordning. Har man redan gjort en sökning med hjälp av sökformuläret står det sökta ordet mitterst i listan. I det röd-svart-guldfärgade fältet finns nederst länkar till var och en av bokstäverna enligt vilka ordbokens uppslagsord är alfabetiserade för att man snabbt ska komma fram i listan. Allra nederst finns en sökruta för att ta fram önskat ord i listan. Ordlistans ord är länkade till ordartiklarna.

Egentliga sökningar i ordboken kan göras på artikelnivå eller som fritextsökningar på exempelnivå. I sökformuläret på skärmen väljer man en av två flikar, den ena för artikelsökning och den andra för exempelsökning. Det är möjligt att göra sökningar med trunkeringstecken. Webbtjänstens instruktioner för sökningar i *Webb-KKS* är mycket ingående, vilket gör att man till fullo kan lära sig utnyttja alla de möjligheter ordboken erbjuder.

3.3.2. Artikelsökning

Vid artikelsökning kan sökelementet vara uppslagsord, ordklassangivelse, förklaring eller finsk översättning, dialektord, socken, källa, bruklighetsmarkör och grammatiska uppgifter. I en del av sökfälten kan sökelementet väljas i en rullgardinsmeny. Detta är en bra hjälp mot onödiga sökmissar på grund av felskrivning.

Min viktigaste invändning gäller det komplicerade i att skriva dialektala sökord och uppslagsord vid artikelsökning. Nu kräver en smidig sökning antingen rätt mycket träning eller förhandskunskaper av användaren. Annars är man för uppslagsordens del hänvisad till sökning i uppslagsordsbalken genom att klicka på begynnelsebokstäverna och sedan skrolla fram till rätt ord.

Uppslagsorden i den karelska ordboken är lemmatiserade i nordkarelsk form och därför är det också motiverat med ett sökfält för dialektord. Att skriva det ord man söker är en utmaning för en icke karelskunnig. Nordkarelskans ortografi innehåller en del tecken som inte går att producera med ett vanligt (skandinaviskt) tangentbord. Därför kan ersättande tecken användas i sökfältet ”uppslagsord”. Bokstäver med diakritiska tecken ersätts med bokstäver utan sådana: i stället för ett *d* med akut accent skrivs i sökfältet *j* eller *t* och i stället för č skrivs *ts*. Däremot kan det också i finsk ortografi förekommande tecknet *š* skrivas i sökfältet. Detta tecken finns dock inte på ett finskt tangentbord. Man måste alltså ta till olika knep för att åstadkomma *š*, t.ex. kopiering från en Word-text.

Lemmatiseringsprinciperna[3] finns förklarade i omtexterna, men trots det kan det bli svårt att göra en sökning på uppslagsordet. En användare som kan finska kan ju alltid ”fuska” lite och gissa sig till en finsk översättning, söka på den och sedan bläddra bland sökresultaten. Alla elektroniska ordböcker kan ju användas i motsatt riktning än vad de ursprungligen är tänkta för, men den metoden ger inte optimalt sökresultat.

Avancerad sökning på en söksträng med avseende på t.ex. socken och bruklighet ger till resultat alla de artiklar där sökelementen förekommer. Detta fungerar dock inte optimalt utan på samma sätt som t.ex. en Google-sökning utan citattecken, dvs. de två sökelementen står inte nödvändigtvis nära varandra. Jag sökte med sökelementen sockennamnet *Oulanka* och brukligheten *itkuvirsissä* (’i gråtkväden’). Sökningen gav nio träffar. En av träffarna, artikeln *tulla*, innehåller ett exempel från Oulanka, men det är inte från ett gråtkväde. Däremot innehåller artikeln ett exempel från ett gråtkväde, men detta är upptecknat i socknen Tulemajärvi. Det återstår alltså för användaren att gallra i sökresultatet för att eventuellt få fram det han eller hon söker. I detta fall finns det inte i materialet några belägg från gråtkväden som skulle vara upptecknade i Oulanka. Det är alltså inte möjligt att söka på två olika sökelement och dessutom lägga in en begränsning på avståndet mellan dem i artikeln. Sådana sökningar kräver att man kommer åt ordbokens databas som text. Ett annat alternativ vore att ordboksprogrammet först plockar ut det man sökt med ett sökelement och sedan möjliggör en sökning till i detta resultat med ett annat sökelement.

3 Eftersom uppslagsorden i ordboken ges i sin nordkarelska form, kan ett dialektord man träffat på i en text och vill slå upp skilja sig från lemmaformen i fråga om diakritiska tecken för suprasegmentala drag, i fråga om vissa konsonanter, speciellt klusiler som i nordkarelskan saknar stämbandston och slutligen i fråga om vissa diftonger.

3.3.3. Exempelsökning

Vid exempelsökning kan sökelementet vara dialektord, förklaring eller finsk översättning, socken, källa, bruklighetsangivelse och grammatiska uppgifter. Nederst i formuläret finns två luckor som kan användas för avgränsning till en viss artikel enligt uppslagsordet och ordklassen. Den senare avgränsningen kan användas t.ex. om man vill skilja mellan adjektiv och adverb. Vid exempelsökning kan trunkeringstecken användas bara då man söker på uppslagsord.

Vid exempelsökning ges sökresultatet i form av en konkordansliknande lista där raderna börjar med uppslagsordet i den artikel där exemplet i fråga finns. Uppslagsordet står inom hakparenteser och är en länk till hela artikeln. Skillnaden till sökresultat i konkordanser är den att söksträngen här inte centreras, däremot är den markerad med gult. Den gula markeringen försvinner dock om man går till artikeln via länken på radens början. Detta är inte riktigt bra, men man kan ju alltid använda webbläsarens egen sökfunktion då man har tagit fram själva artikeln.

En exempelsökning med avgränsning till uppslagsord kan kännas irrelevant, men i fråga om mycket långa artiklar är en sådan sökning användbar. T.ex. i artikeln *tulla* (’komma’) ingår 338 exempel. Det är lättare att få en överblick över exemplen då de är uppställda som i en konkordans. Om man bland dem vill söka de exempel som är upptecknade i Tulemajärvi, gör man en avancerad sökning på sökelementen uppslagsord och socken kombinerade och får fram artikelns alla sex exempel som är upptecknade just i Tulemajärvi. Här går det alltså att söka på det sätt som efterlystes ovan.

4. Vad kunde ytterligare önskas?

Då man tar fram elektroniska ordböcker, framför allt webbordböcker kan man praktiskt taget strunta i kravet på att spara utrym-

me. Den elektroniska ordbokens fördelar kommer på många sätt fram då man använder *Webb-KKS*. En egenskap hos elektroniska ordböcker har dock inte utnyttjas och det är möjligheten att visa respektive dölja element i ordboksartiklarna.

Ett sökresultat kan vara en mycket lång artikel som kan vara besvärlig att läsa då man har hela texten på skärmen. Genom att visa respektive dölja element får användaren fram det han eller hon verkligen är intresserad av. Den som bara är intresserad av t.ex. ett ords betydelser och kanske de finska översättningarna, kunde tänkas vilja gömma exemplen och vice versa. Utgående från databasens nuvarande kodning är detta troligen helt möjligt. Kanske en idé för det fortsatta utvecklingsarbetet med användargränssnittet? Så kunde denna webbordbok också bli en ”andra generationens” ordbok. Ett gott exempel på en sådan är *Oxford English Dictionarys* webbversion *OED Online*.

5. Avslutning

Webb-KKS är en ”första generationens” webbordbok och är som sådan jämförbar med *Svenska Akademiens Ordbok* på Internet. Ett liknande utvecklingssamarbete mellan lexikografer, dataplanerare och programmerare ligger troligen bakom de båda storverken. Samarbetet har burit en vacker frukt i och med att KKS nu finns gratis tillgänglig för forskare och den stora allmänheten. Detta ökar demokratin inom kunskapens område i och med att ordboken kan användas också där det inte finns tillgång till den tryckta ordboken. Utgivningen på Internet garanterar också en internationell spridning. Många av den karelska ordbokens användare finns t.ex. i Ryssland.

Informationen i det kodade manuskriptet till ordboken ser ut att vara så gott som maximalt utnyttjad. Arbetet är så lyckat att den i dessa sammanhang än så länge något oortodoxa tanken

infinner sig att den tryckta ordboken förlorar en hel del i jämförelsen med webbordboken. Å andra sidan hade det naturligtvis varit synd att avbryta utgivningen av en bokserie som redan funnit sin väg till universitet, bibliotek och också till ett antal privata bokhyllor. Diskussionen omkring publiceringen av några andra omfattande ordböcker som fortfarande är under arbete vid Forskningscentralen för de inhemska språken pågår dock som bäst. I framtiden kommer den elektroniska utgivningen att vara självklar. Jag ser fram emot att se också andra av forskningscentralens ordböcker på Internet, helst så att de alla kan konsulteras med hjälp av samma sökformulär.

Problemet med de svårskrivna sökorden står på arbetslistan då man vid Forskningscentralen för de inhemska språken utvecklar användargränssnittet för *Webb-KKS*. En del förbättringar har redan gjorts under våren 2010 medan detta skrivs. Vid exempelsökning på dialektord kan man nu i stället för en del tecken i den fonetiska skrift som används i ordboken skriva tecken som kan åstadkommas på ett vanligt tangentbord. T.ex. tecknet för ett muljerat *d* (d + akut accent) kan i sökningarna skrivas d' (d + vanligt accenttecken). Avsikten är att åstadkomma ett hjälpfönster där man kunde klicka fram specialtecken och bokstäver försedda med diakritiska tecken. Det kunde dessutom vara bra med ett extrafönster med en lista över omskrivningarna av vissa bokstäver och bokstavskombinationer vid lemmatiseringen. Extrafönstret kunde t.o.m. vara optionellt så att man kan klicka bort det om man tycker att man inte behöver det.

Men för *Webb-KKS* (och för *SAOB*) återstår ännu att ta steget fullt ut till att vara en "andra generationens" webbordbok där delar av omfattande ordboksartiklar kan lyftas fram beroende på vad man är intresserad av.

Till sist vill jag kommentera webbordbokens namn. Varför kalla webbordboken för en version? Ordböcker utarbetas numera som databaser som sedan kan ges ut på olika plattformer beroende

bl.a. på hur man vill presentera databasens innehåll och vilka möjligheter till konsultering av ordboken (slå upp vs. söka) man vill erbjuda sina användare. Arbetsgången med *Webb-KKS* har varit något mera komplicerad än detta och visst har den tryckta ordboken varit primär eftersom elektroniska ordböcker knappast var i tankarna då ordbokens första band kom ut 1968. Men i dagens läge är både den tryckta och den elektroniska ordboken ”versioner” och då kanske vore förslagsvis namnet *Karjalan kielen verkkosanakirja* (sv. Webbordbok över karelskan) bättre motiverad.[4] På detta sätt tar man inte ställning till huruvida någondera av de två utgivningsformerna är primär eller sekundär. *Webb-KKS* förtjänar i alla fall på inget sätt att bli betraktad som en sekundär version.

Litteratur

Ordboken

Karjalan kielen sanakirja 1–6. 1968–2005. Utgivare: Finsk-Ugriska Sällskapet: Lexica Societatis Fenno-Ugricae 16 band 1–6, Forskningscentralen för de inhemska språken: Kotimaisten kielten tutkimuskeskuksen julkaisuja 25. Helsinki 1–3, Vammala 4–6.

Karjalan kielen sanakirjan verkkoversio (Webbversionen av Ordbok över karelskan). http://kaino.kotus.fi/cgi-bin/kks/kks_etusivu.sgi (mars, april 2010)

4 Efter att detta skrivits har *Webb-KKS* uppdaterats den 7.9.2010. Redaktionen har fått möjlighet att ta del av skribentens kommentarer och en del av dem har beaktats. Den mest synliga ändringen är den att webbordboken nu faktiskt heter *Karjalan kielen verkkosanakirja*.

Övrig litteratur

Tanner, Satu 2006: En dokumentation av det karelska språket. I: *LexicoNordica* 13, 169–183.

Internethänvisningar

Kaino http://kaino.kotus.fi/ (mars, april 2010)

Miikkulainen, Raisa 2005: *Karjalan kieli innostaa* (Det karelska språket inspirerar). Intervju på Forskningscentralens för de inhemska språken webbplats http://www.kotus.fi/index.phtml?s=291 (mars 2010)

Oxford English Dictionary (OED) Online http://dictionary.oed.com/

Svenska Akademiens ordbok (SAOB) http://g3.spraakdata.gu.se/saob/

Ilse Cantell
forskare, fil. lic.
Rödbergsgatan 23 I 197
FI-00150 Helsingfors
il.cantell@pp.inet.fi

Lexicography in the 21st Century

Cathrine Fabricius-Hansen

Lexicography in the 21st Century. In honour of Henning Bergenholtz, udg. af Sandro Nielsen og Sven Tarp. John Benjamins Publishing Company, Amsterdam/Philadelphia 2009.

1. Indledning

Det foreliggende værk er udgivet i anledning af at professor Henning Bergenholtz, leder af *Center for Leksikografi – Forskning i behovstilpasset informations- og datatilgang* ved Handelshøjskolen, Århus Universitet, er fyldt 65 (i 2009). Det indeholder 14 artikler (i alt 308 sider) forfattet af fagfæller i ind- og udland (repræsenteret ved Sverige, Norge, Island, Tyskland, Spanien, Canada og Sydafrika). I tillæg har udgiverne skrevet en kort indledning og udarbejdet et 25 siders "bibliovita" i form af en kronologisk liste over de vigtigste publikationer (ordbøger og teoretiske bidrag) Henning Bergenholtz er med- eller eneforfatter af. Antologien genspejler i sin helhed både Henning Bergenholtz' internationale forskningsprofil og den store betydning han har haft for leksikografisk teori og praksis, og kan i så henseende siges at have opfyldt et af sine formål på udmærket vis.

De 14 artikler er nogenlunde ligeligt fordelt på 5 temaområder (se afsnit 2). De er i det store og hele orienteret mod fagspecifik leksikografi ("lexicography for special purposes") og gennemgående forankrede i "functional lexicography", dvs. erkendelsen af at en ordbogs funktion – om den f.eks. skal bruges aktivt eller passivt i sproglig kommunikation eller mere som et middel til øget erkendelse – har afgørende betydning for hvilke oplysninger den bør

indeholde og hvordan den bør struktureres. Jeg vil nedenfor først give en kort kommenteret indholdsoversigt (afsnit 2) og derefter en sammenfattende vurdering ud fra mit eget ståsted, som vel at mærke ikke er leksikografens eller leksikologens, men den almene lingvists og fremmedsprogslærerens (se afsnit 3).

2. Kommenteret indholdsoversigt

Del I med titlen *The dictionary, dictionary structures and access routes* omfatter fire artikler om ordbøger, deres opbygning og hvordan man henter oplysningerne ud af dem ("access routes"). Den første – "Sinous lemma files in printed dictionaries: Access and lexicographic functions" af Rufus H. Gouws – giver en nyttig oversigt over forskellige typer horisontal anordning af lemmaer inden for en overordnet vertikal alfabetisk makrostruktur, herunder såkaldt rededannelse og nichedannelse. Hovedbudskabet er at anvendelsen af sådanne virkemidler må tilpasses brugerens behov i højere grad end hvad man almindeligvis finder i trykte ordbøger, og at det er specielt vigtigt at sikre en effektiv tilgang også til lemmaer der bryder med den typiske vertikale alfabetiske præsentation. Dataadgang – eller tilgængelighed – er hovedtemaet for det tredje bidrag, "Reflections on data access in lexicographic works" af Sven Tarp. Det anskueliggør hvilke praktiske konsekvenser hensynet til brugeren og hendes/hans behov i varierende situationer kan eller bør have for tilrettelægningen – og hvilken fleksibilitet ny teknologi giver mulighed for i så henseende. I artikel nr. 2, "Reviewing printed and electronic dictionaries: A theoretical and practical framework", diskuterer Sandro Nielsen forskellige typer evaluering af ordbøger og hvilke krav de bør opfylde. Det er svært at erklære sig uenig i hans synspunkter. Noget andet er om kriterierne for "god" kritik er realistiske. Artikel nr. 4, "Hybrid text constituent structures of dictionary articles: A contribution

to the expansion of the theory og textual dictionary structures", er en ekstremt abstrakt metaleksikografisk analyse af hvilke slags tekstelementer ordbogsartikler kan indeholde. Den er skrevet af Herbert Ernst Wiegand, en nestor i moderne leksikografi, og udmærker sig ved udelukkende at referere til arbejder han selv er (med)forfatter af.

Del II bærer titlen *Dictionary functions and users* og indeholder tre bidrag med temmelig forskellig tematik. I artikel nr. 5, "On production-oriented information in Swedish monolingual defining dictionaries", orienterer Sven-Göran Malmgren om hvordan *Svensk ordbok* (1986) og dens efterfølgere håndterer information som er relevant ved aktiv brug af sproget. Det sker med reference til de funktionsbaserede kriterier Henning Bergenholtz og Vibeke Vrang (2004, 2005, 2006) anvender i deres indflydelsesrige vurderinger af *Den Danske Ordbog* (2003-2005). Det efterfølgende bidrag (artikel nr. 6) – "Balancing the tools: The functional transformation of lexicographic tools for tourists" af Patrick Leroyer – skitserer fremtidens leksikografiske hjælpemidler for turister – hjælpemidler som i hans visioner ligger fjernt fra de traditionelle rejseordbøger. I den tredje artikel (nr. 7), "Lexicography and language planning in Scandinavia and the Netherlands", beskriver Lars S. Vikør med udgangspunkt i en artikel af Bergenholtz og Gouws (2006) hvordan ordbøger kan bruges og bliver brugt som sprogpolitisk redskab i Norge.

Del III, *Subject-field classification and introductions*, indeholder et bidrag (artikel nr. 8) af Bo Svensén ("Subject-field classification for metalexicography") om et emne som er centralt i fagspecifik leksikografi, nemlig klassificering af relevante emneområder. Desværre er det område forfatteren har valgt ud, nemlig metaleksikografi, næppe egnet til at vække interesse ud over leksikografernes egne rækker. Den efterfølgende artikel (nr. 9) af Pedro A. Fuertes-Olivera, "Systematic introductions in specialised dictionaries: Some proposals in relation to accounting dictionaries", giver en

instruktiv, konkret indføring i et andet centralt emne, nemlig ordbogsindledninger, med udgangspunkt i det pionerarbejde Bergenholtz og Tarp (1995) har gjort på området.

Del IV drejer sig om *Data retrieval and corpus lexicography* – begge dele højst aktuelle temaer i moderne (elektronisk) leksikografi. F. J. Prinsloo diskuterer i artiklen "The role of corpora in future dictionaries" (artikel nr. 10) centrale spørgsmål som korpusdesign, annotering, lemmatisering og nytten af leksikalske resurser som FrameNet og WordNet. Det andet bidrag (artikel nr. 11) i denne del, "Lexicographical data in natural-language systems" af Franziskus Geeb, omhandler muligheden for og den mulige nytteværdi af at bruge leksikografiske data i interaktive chatbox-systemer. Begge artikler er – rimeligt nok – relativt tekniske, men åbner samtidig spændende perspektiver.

Del IV – *Collocations and phraseology* – indeholder tre artikler. De to første – "A methodology for describing collocations in a specialised dictionary" (nr. 12) af Marie-Claude L'Homme og "Lexicographic description: An onomasiological approach on the basis of phraseology" (nr. 13) af Jón Hilmar Jónsson – præsenterer konkrete projekter af almen metodisk interesse. Den tredje (og sidste i hele samlingen), "Item-specific syntagmatic relations in dictionaries" (nr. 14) af Thomas Herbst, diskuterer med eksempler fra engelsk ordbogstradition to vigtige udfordringer for ordbøger som (også) skal kunne bruges af fremmedsprogslørnere i aktiv sammenhæng, nemlig koderingen af syntaktiske oplysninger (rektion, valens) og spørgsmålet om hvordan kollokationer og "item-specific" konstruktioner bedst kan præsenteres.

3. Sammenfattende vurdering

Som nævnt i afsnit 1 er denne anmelder ikke leksikograf – og derfor heller ikke den rette til at give en fagfællevurdering af den fore-

liggende antologi. Jeg har nærmet mig opgaven som interesseret lingvist og ordbogsbruger med erfaring fra fremmedsprogsundervisning (tysk) – og lært en del nyt både om eksisterende ordbøger og om leksikografiens udviklingsmuligheder i det 21. århundrede. Samtidig har jeg fået et indtryk af hvad leksikografisk forskning er optaget af, selv om jeg naturligvis ikke kan afgøre om det foreliggende festskrift er helt repræsentativt i så henseende.

Antologien henvender sig til leksikografer i vid forstand (med god grund i betragtning af anledningen), dvs. både forskere i leksikografi og praktiske leksikografer. Imidlertid har overraskende mange af bidragene ikke sidstnævnte som deres primære målgruppe, hvilket bidrager til at give bogen en vis akademisk slagside. Det kan f.eks. være svært at forestille sig hvordan de ideelle mål der skitseres i bidragene om evaluering af ordbøger (artikel nr. 2) og klassificering af emneområder (nr. 8), kan omsættes i praksis. Man kan også spørge sig om fremtidens turister faktisk vil(le) benytte sig af det elektroniske værktøj Leroyer (artikel nr. 6) forestiller sig, dvs. om nytteværdien står i et rimeligt forhold til hvad det vil(le) koste at udvikle et sådant værktøj. Wiegands bidrag (artikel nr. 4) er et markant eksempel på leksikografisk forskning som næppe kan have den store interesse for leksikografisk praksis.

Som tilfældet er med alle antologier, uanset hvor tæt sammenhæng udgiverne måtte ønske sig mellem bidragene, er der også i dette bind artikler der nok ligger lidt på siden af det man umiddelbart venter sig på baggrund af bogens titel. Det gælder artiklerne om ordbøger og sprogplanlægning (nr. 7) og brug af leksikografiske data i chatbox-systemer (nr. 11), som på den anden side giver et nyt blik på også leksikografiens brugsområder. Andre bidrag (5, 12 og 13) er tematisk set klart relevante, men så specifikke at de har svært ved at vække interessen hos en læser uden for fagkredsen.

Temaet korpusleksikografi, som får en informativ, men forholdsvis teknisk behandling i artikel nr. 10, er indlysende centralt

i det 21. århundredes leksikografi og kunne med fordel have været taget op i flere bidrag. Det havde også været nærliggende at nævne det store tyske elektroniske ordbogsprojekt *Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts* (DWDS) som eksempel på en fremtidsrettet og brugervenlig sammenkobling af ordbøger og korpora.

Personligt har jeg haft mest udbytte af de artikler – nr. 1, 3, 9 og 14 – der vedrører problemer jeg har mærket på min egen krop som ordbogsbruger, og som jeg har set studenter slås med, dvs. problemer med at finde frem til de oplysninger man er interesseret i og forstå de oplysninger der gives f.eks. af syntaktisk karakter uden at måtte konsultere en liste med uigennemsigtige forkortelser. Det foreliggende bind giver et klart indtryk af at der på disse områder er sket og fortsat vil ske meget – og at fremtidens ordbøger vil være meget forskellige fra dem min egen generation er vokset op med. Det skyldes dels den principielle drejning mod funktionel leksikografi, hvor ordbogsbrugeren og de forskellige formål ordbøger kan bruges til, står centralt, dels naturligvis anvendelsen af moderne teknologi (inklusive elektroniske korpora) i leksikografisk sammenhæng. Og der er ingen tvivl om at den person der æres med dette værk, har en del af ansvaret for denne udvikling – og således æres med rette.

Litteratur

Ordbøger

Den Danske Ordbog. Bind 1-6. København: Gyldendal 2003-2005.
Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS). http://www.dwds.de/woerterbuch
Svensk ordbok. Stockholm: Esselte studium. 1986.

Anden litteratur

Bergenholtz, Henning og Rufus H. Gouws 2006: How to do Language Policy with Dictionaries? I: *Lexikos* 16, 13-45.

Bergenholtz, Henning og Sven Tarp (udg.) 1995: *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: John Benjamins.

Bergenholtz, Henning og Vibeke Vrang 2004: Ny dansk ordbog i seks bind for sekretærer og forskere. I: *LexicoNordica* 11, 165-189.

Bergenholtz, Henning og Vibeke Vrang 2005: Den Danske Ordbog bind 2 (E-H) og 3 (I-L) – en ordbog for folket eller for akademikere? I: *LexicoNordica* 12, 169-187.

Bergenholtz, Henning og Vibeke Vrang 2006: Den Danske Ordbog: en ordbog for lingvister! I: *LexicoNordica* 13, 185-196.

Cathrine Fabricius-Hansen
professor
ILOS, Universitetet i Oslo
Postboks 1003, Blindern
NO-0315 Oslo
c.f.hansen@ilos.uio.no

Värd ett besök – om DSL:s nya webbsida ordnet.dk

Ruth Vatvedt Fjeld & Sven-Göran Malmgren

Elektroniska lexikon blir allt vanligare, både i och ännu mer utanför Norden. Många frågar sig om den traditionella pappersordboken har någon framtid, när den ena utmärkta ordboken efter den andra läggs ut gratis på nätet, med eller utan extra sökmöjligheter. I Sverige har det största ordboksförlaget, Norstedts, efter lång tvekan till sist beslutat sig för att lägga ut många av sina en- och tvåspråkiga ordböcker, tills vidare enbart med reklamfinansiering (www.ord.se). För mer vetenskapligt intresserade ordboksanvändare finns sedan länge den oumbärliga SAOB (Svenska Akademiens ordbok) gratis på nätet (http:/g3.spraakdata.gu.se/saob/), med utmärkta sökmöjligheter. Vidare pågår arbete med sikte på att få upp den nya definitionsordboken SO (Svensk ordbok utgiven av Svenska Akademien, 2009, recenserad av Kristina Nikula i detta nummer av LexicoNordica) på nätet.

I Norge finns en samling lexikografiska uppslagsverk på nätet under Kunnskapsforlagets webbsida *ordnett.no*, som anmäldes i förrförra numret av LexicoNordica (Ims 2008). Webbsidan är inte minst ett mycket bra översättningshjälpmedel, då de flesta av dess ordböcker är tvåspråkiga, men den innehåller också några enspråkiga ordböcker. Dessa är dock inte kopplade till databaser och de är därför inte särskilt sökbara. Efterhand är det meningen att den stora Riksmålsordboken ska tillkomma, och när det har skett kommer det åtminstone att finnas en större enspråkig norsk elektronisk ordbok på nätet. Redan nu finns dock (utanför den nämnda webbsidan) Bokmålsordboka och Nynorskordboka på nätet, med ganska goda sökmöjligheter. När Norsk ordbok (i 14

band) blir färdig 2014, blir det spännande att se hur den så småningom kommer att ta sig ut som nätordbok.

Men redan nu har danska ordboksvänner fått en gåva som överträffar allt som hittills skådats i Norden, den lexikografiska webbsidan *ordnet.dk*. Den är utarbetad på Det Danske Sprog- og Litteraturselskab (DSL) och inrymmer tre (egentligen fyra) huvudkomponenter, som framgår av portalsidan (se figur 1): den modernspråkliga Den Danske Ordbog (DDO 1–6, 2003–2005), den historiska Ordbog over det danske Sprog (ODS 1–28, 1918–56) samt en korpus på 56 miljoner ord över danskan från 1990 och framåt. Dessutom – det förklarar sajtens namn – kan man via artiklarna i DDO komma till en version av det danska WordNet (DanNet).

Figur 1: Portalen till ordnet.dk

Ordnet.dk är också av stort intresse för användare utanför Danmark, eller rättare sagt icke-danskar. För det första borde de datatekniska lösningarna, den grafiska designen m.m., kunna inspirera skapare av ordbokssajter utanför Danmark. För det andra är webbsidan ett utomordentligt hjälpmedel för danskstuderande (med annat modersmål än danska), säkert tiotusentals personer.

Och slutligen finns ett antal skandinavister – både lexikologer och grammatiker – som ägnar sig åt kontrastiva studier, t.ex. mellan danska och svenska eller mellan danska och norska, och som hittills ständigt använt nordiska pappersordböcker. I det sammanhanget är ordnet.dk en fantastisk resurs.

Vi ska se på webbsidans tre huvudkomponenter var för sig och även på möjligheterna att navigera mellan ordböckerna och korpusen. Vi börjar med korpusen, kallad KorpusDK.

1. KorpusDK

Korpusen är såvitt vi förstår identisk med den korpus (Korpus 2000) som länge har varit tillgänglig på nätet, och som en av oss tidigare presenterat för icke-danska användare (Malmgren 2003). Vi har använt den korpusen i åtskilliga år och blivit vana vid den, och det ska erkännas att vi hade vissa begynnelsesvårigheter i umgänget med den nya versionen av korpusen. Korpusen innehåller dels texter från ca 1990, dels texter från ca 2000, sammanlagt ca 56 miljoner ord. En delmängd av korpusen användes som underlag för DDO. Det utlovas att korpusen efter hand kommer att uppdateras med nyare texter.

Poängen med den nya versionen av korpusen är framför allt de förbättrade sökmöjligheterna. Vi ska se på några av dem, men även på sådana sökmöjligheter som redan fanns i Korpus 2000.

För det första kan man självklart söka på lemman och inte bara på ordformer. Eller rättare sagt, om man skriver in söksträngen *springer*, så får man en fråga om man vill ha textexempel enbart på presensformen eller om man vill ha alla former av verbet *springe*. I det senare fallet får man exempel med *springe*, *springer*, *sprang*, *sprunget* etc. Detta är numera inte så märkligt. De norska lexikografiska korpusarna har denna funktion, men det är fortfarande (april 2010) inte möjligt att göra sådana lemmatiserade sökningar

i den svenska Språkbanken (http://spraakbanken.gu.se), som annars har många förtjänster.

För det andra kan man till ett givet ord (t.ex. *modstand*) få fram s.k. naboord, ord som är statistiskt överrepresenterade i närheten av det givna ordet (i det här fallet t.ex. *yde*, *passiv*). Många sådana naboord bildar äkta kollokationer (icke-triviala fraser) med huvudordet, varför denna sökmöjlighet innebär ett utomordentligt lexikografiskt hjälpmedel. En liten svaghet är dock att man inte får naboorden lemmatiserade. Om man vill se vilket verb som är vanligast i kombination med *modstand*, så finner man det inte direkt utan måste lägga ihop frekvenserna för *yder*, *yde* etc. På den punkten har KorpusDK möjlighet att raffinera sina resultat, då lemmatisering av det här slaget låter sig göra förhållandevis lätt.

Men naboordsprogrammet är redan nu ett mycket användbart hjälpmedel, utom för lexikografer t.ex. för översättare och alla som vill skriva god danska. Vid ord som *modstand* kan man t.ex. ofta vara osäker om vilket som är det mest brukliga förstärkande adjektivet – t.ex. *voldsom*, *stærk*, *hård* eller *kraftig*. I sådana fall ger KorpusDK ofta god vägledning.

För det tredje – och det är nog den viktigaste förbättringen – kan man söka på grammatiska mönster i korpusen (under rubriken ”Formel søgning”). Det är visserligen en smula komplicerat; man behöver en s.k. taggtabell med koder för olika morfologiska kategorier, t.ex. för substantiv i bestämd form plural. Man behöver dock inte själv skriva in koderna utan kan kopiera dem från taggtabellen. Notationen som används påminner om notationen i Parole-korpusen i den svenska Språkbanken, där man länge har kunnat söka på grammatiska kategorier, med någorlunda hygglig disambiguering eller träffsäkerhet; jfr nedan.

Det ser dock ut som om listan över morfologiska kategorier inte är riktigt lika fullständig i KorpusDK som i Parole-korpusen; exempelvis kan vi inte hitta kategorin infinitiv. Dessutom är instruktionerna till användarna inte tillräckliga. Man måste komma

ihåg att det är fråga om rätt komplicerade sökningar också för vana lingvister. De exempel som ges domineras av huvudsakligen lexikala sökningar, söksträngar med en följd av ordformer eller lemman. Det finns inte (april 2010) ett enda exempel på en grammatisk sökning som innefattar kategorin ”morph”, dvs. sökningar på en bestämd böjningsform. Antag t.ex. att vi vill veta vilka verb som förekommer som andra verb i fraser av typen *komme springende*. Vi måste då skriva in den söksträng som motsvarar (lemmat) *komme* + godtyckligt verb i presens particip. På grundval av instruktionen går det inte att räkna ut hur det går till, även om vi till sist lyckades lista ut hur man gör. Instruktionen måste följaktligen förbättras.

Erfarenheterna från Sverige tyder på att folk i allmänhet knappast tar sig tid att göra så pass avancerade sökningar i Parolekorpusen (jfr också Trap-Jensen 2010). Men för forskare och studerande på lite högre nivåer är det fråga om utmärkta hjälpmedel, och det är vackert så.

Vid både lexikala och grammatiska sökningar i korpusar är homografi en störande faktor. Man kanske vill ha textexempel på substantivet *skade*, men då upptäcker man snart att de är uppblandade med exempel på **verbet** *skade*. En fjärde viktig poäng med KorpusDK är att homografa ord disambigueras med hygglig säkerhet (den kan aldrig bli hundraprocentig). Vi ska se på träffsäkerheten i KorpusDK när det gäller två homografa ord, eller egentligen fyra ord, *skade* (verb och substantiv) och *cykler* (verb och substantiv). Vi anger antal rätt på de första hundra konkordansraderna när man söker på *skade*/verb, *skade*/substantiv, *cykler*/verb och *cykler*/substantiv i KorpusDK och på motsvarande ord med samma homografi i den svenska Parole-korpusen (*skada*/*cyklar*) och i den norska bokmålskorpusen LBK (*skade*/*sykler*).

	skade/ skada v	skade/ska-da s	cykler/ sykler/ cyklar v	cykler/ sykler/ cyklar s
KorpusDK	72	100	88	99
Parole	100	76	98	74
LBK	92	99	79	92

Tabell 1: Procent rätta angivelser vid homografa ord i KorpusDK, Parole och LBK

Sammantaget blir andelen korrekta träffar i KorpusDK i detta (mycket lilla) test 90 procent, i Parole-korpusen 87 procent och i LBK-korpusen 90,5 procent.

Vi ser alltså att Parole-korpusen, som gjordes på 90-talet, nästan håller jämna steg med KorpusDK, och att LBK:s resultat t.o.m. är aningen bättre än KorpusDK:s. Vad KorpusDK:s prestanda beträffar, är de hyggliga, men inte påfallande bra. Det brukar sägas att 95 procent rätt är ett bra resultat, och dit är det fortfarande en bit kvar för KorpusDK, åtminstone om man får döma av de båda aktuella exemplen. Intressant är att träffsäkerheten i KorpusDK och LBK är mycket god när man söker på substantiven *skade* och *cykler* men betydligt sämre när man söker på verben. Samma skillnad, men så att säga omvänt, gäller för Parole-korpusen. Det kan vara så att disambigueringsprogrammet i svåra fall prioriterar substantivalternativet i KorpusDK och LBK och verbalternativet i Parole-korpusen.

Från användarnas synpunkt finns enligt vår mening en liten försämring, just när det gäller disambiguering, i KorpusDK jämfört med Korpus2000. Om man sökte på *cykler* i Korpus 2000, fick man snabbt en fråga om man menade substantivet eller verbet, och det var bara att klicka i en liten cirkel för det önskade alternativet. I KorpusDK är det, åtminstone vad vi kunnat finna, betydligt mer komplicerat. Det finns i princip två sätt, båda ganska obekväma. En av möjligheterna är att gå till den nämnda taggtabellen och skriva word=”skade” & pos = ”N.*” (f.ö. precis som i Parolekorpusen).

Men det är en detalj. KorpusDK är en verkligt förnämlig korpus, som blir ännu förnämligare genom att den är kopplad till två lika förnämliga ordböcker. Vi återkommer snart till det.

2. ODS

Det är oerhört värdefullt att den stora danska historiska ordboken har kommit upp på nätet. Den spännande historien om hur det gick till när ODS digitaliserades har berättats flera gånger (se t.ex. Bojsen & Trap-Jensen 2005). Facit är att en nästan korrekt nätversion av ODS är gratis tillgänglig på nätet. Ännu finns dock inga nämnvärda sökmöjligheter, med ett viktigt undantag: man kan göra trunkeringar på uppslagsord, t.ex. ta fram alla uppslagsord som slutar på *-barn*. Ännu är inte heller ODS-supplementet integrerat med den gamla ordboken, men allt detta kommer säkert så småningom. Just i det här sammanhanget finns det därför inte så mycket att säga om ODS-komponenten i ordnet.dk, men vi kan göra en kort jämförelse med Svenska Akademiens ordboks (SAOB:s) nätversion (se t.ex. Cederholm & Rogström 2000).

Det är givet att sökmöjligheterna i ODS än så länge är vida sämre än i SAOB. Men kanske kommer man om några år att kunna göra sökningar av typen ”engelska lånord i danskan under 1800-talet” – och i så fall kan man vara säker på att man får ett helt korrekt resultat, vilket tyvärr inte är fallet med SAOB. En uppenbar fördel med ODS är att den sedan länge är färdig. Ytterligare en finess med nätversionen av ODS är att den är försedd med en bibliografi över alla vetenskapliga artiklar om ordboken, från början av 1900-talet och framåt. Det är en idé som det vore bra om SAOB:s nätsida tog efter. Det behöver inte kosta så mycket möda eftersom största delen av denna bibliografi finns förtecknad i Lundbladh (2003).

3. DDO (inklusive DanNet)

De många förtjänsterna hos pappersversionen av DDO har framhållits i åtskilliga recensioner, t.ex. Pedersen (2004), Malmgren (2004) och även Bergenholtz/Vrang (2004, 2005). De behöver inte upprepas här eftersom de givetvis har följt med till nätversionen. Här ska vi framför allt diskutera vad nätversionen ger utöver pappersversionen. Vi börjar med att se på DDO-komponentens ingångssida (se figur 2).

Figur 2: Ingångssidan till DDO-komponenten i ordnet.dk

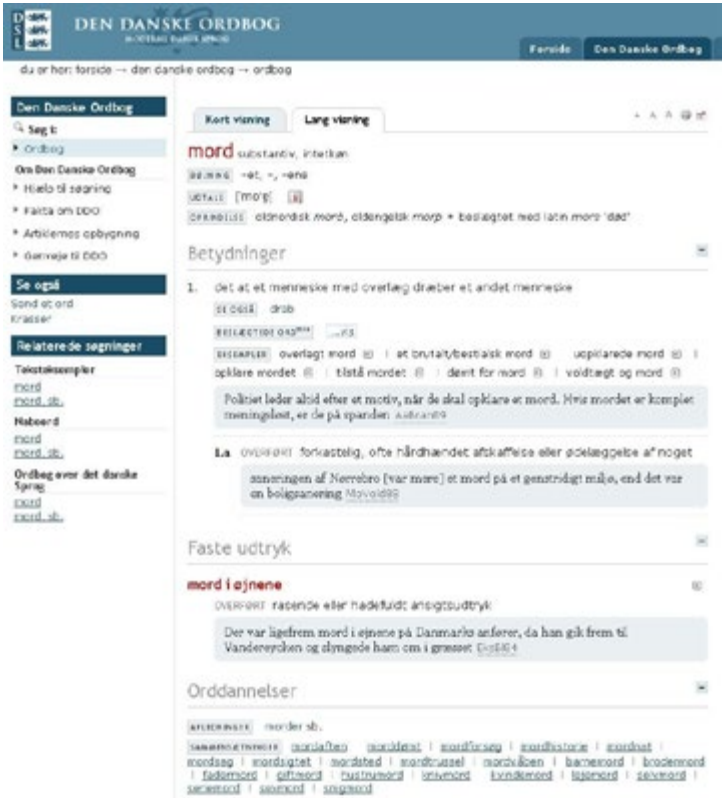
Ingångssidan gör ett mycket trevligt intryck. I centrum ser man omedelbart sökfönstret och i vänstermarginalen finns länkar till instruktioner och till mer information om DDO. Det är också fint att man så tydligt vänder sig till användarna med feedback och att de får figurera på sidan med sina egna (partiella) namn. Att man lyfter fram ett nytt ord varje dag är också något som bör stimulera intresset hos användarna.

En iögonenfallande skillnad mellan DDO:s nät- och pappersversion är att nätversionen innehåller nära 100 000 uppslagsord mot pappersversionens 63 000. Till någon del är det troligen fråga om ord som uppträtt i danskan efter 2003, men det ojämförligt största tillskottet utgörs av sammansättningar och avledningar som togs upp som exempel utan att förklaras i pappersversionen. I en olycklig förlagsreklam 2003 angavs antalet uppslagsord i DDO till 100 000, något som påtalades av Bergenholtz/Vrang (2004). Nu kan DDO alltså helt ärligt ståta med siffran 100 000 uppslagsord, men trots allt måste just den förändringen till stor del sägas vara av kosmetiskt slag. Att sammansättningsexemplet *mordaften* lyfts upp till lemmat **mordaften** (med böjning och uttal, men ingen, eller i andra fall helt trivial, betydelsebeskrivning) innebär inget stort informationstillskott. Med någon tillspetsning kan man säga att förändringen består i att ord i mager stil har avancerat till ord i fetstil – och därmed, med en rent formalistisk definition av termen *lemma*, blivit lemman.

Låt oss nu se på en artikel i DDO:s nätversion, nämligen artikeln **mord**; se figur 3. (Av utrymmesskäl tvingas vi välja en ganska kort artikel; i längre artiklar framträder DDO:s starka sidor ännu tydligare.) Vid genomgången bör det hållas i minnet att vi är medvetna om att det inte finns någon allmän nätordbok vare sig i Danmark, Norge eller Sverige som kvalitetsmässigt kan jämföras med DDO; vi tillåter oss ändå några kritiska synpunkter som ska ses som förslag om hur en utomordentlig nätordbok kan bli ännu bättre.

Strax ovanför lemmatecknet **mord** ser vi två flikar med texten ”Kort visning” resp. ”Lang visning”, något som säkert uppskattas av många läsare. Den version som bilden visar och som vi diskuterar här är den långa.

Vi ser först på den formella delen av artikeln och fastnar till en början för uttalet, där vi konstaterar att ordboken (ännu) inte är talande, något som skulle innebära en stor förbättring. Visserligen kan man genom en klickning komma från den fonetiska beskriv-

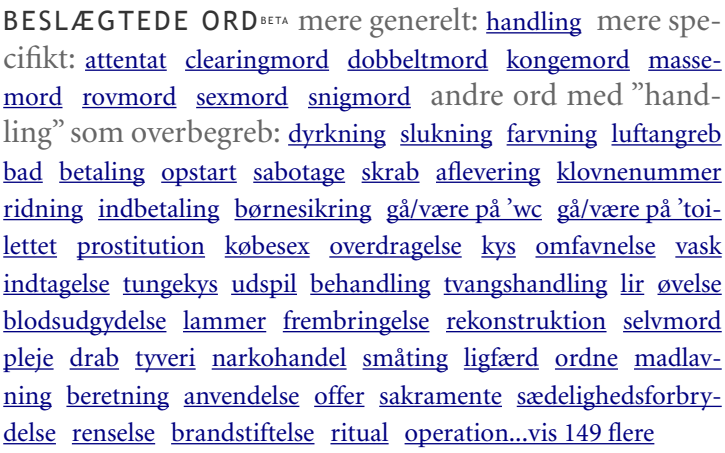


Figur 3: Artikeln **mord** i nätversionen av DDO (bilden är något beskuren från höger)

ningen till förklaringar av notationen, men den IPA-inspirerade uttalsnotationen i DDO är nog trots det ganska svår för många användare. I samband med böjningsuppgifterna (strax före uttalsuppgifterna) kan vi konstatera, att alla eller nästan alla böjningsformer är sökbara liksom i nätversionerna av Bokmålsordboka och Nynorskordboka, men i motsats till vad som är fallet i svenska nätordböcker.[1]

1 Det verkar inte gå att hitta fram till grundformen från bestämd form superlativ av adjektiv, t.ex. *hyppigste*.

Efter definitionen och halvsynonymen *drab* (som är klickbar, man kan genast komma över till den artikeln) kommer vi till en av de viktigaste nyheterna i nätversionen, anknytningen till det danska ordnätet. Man klickar under ”BESLÆGTEDE ORD” på ”vis” och får se över 100 mer eller mindre besläktade substantiv (se figur 4). Det bör nämnas, att ordnätet också finns tillgängligt i grafisk form, samt att det som i skrivande stund ligger ute än så länge är en betaversion.


BESLÆGTEDE ORD BETA mere generelt: handling mere specifikt: attentat clearingmord dobbeltmord kongemord massemord rovmord sexmord snigmord andre ord med ”handling” som overbegreb: dyrkning slukning farvning luftangreb bad betaling opstart sabotage skrab aflevering klovnenummer ridning indbetaling børnesikring gå/være på 'wc gå/være på 'toilettet prostitution købesex overdragelse kys omfavnelse vask indtagelse tungekys udspil behandling tvangshandling lir øvelse blodsudgydelse lammer frembringelse rekonstruktion selvmord pleje drab tyveri narkohandel småting ligfærd ordne madlavning beretning anvendelse offer sakramente sædelighedsforbrydelse renselse brandstiftelse ritual operation...vis 149 flere

Figur 4: Resultat av sökning på ”besläktade ord” i artikeln **mord** i DDO

Dessa uppslag är ordnade så att man först kommer till ”mere generelt”, alltså överbegreppet eller hyperonymen, i detta fall *handling*, och sedan till ”mere specifikt”, här t.ex. *attentat*, *clearingmord* och *dobbeltmord*. Om man inte vet t.ex. vad *clearingmord* är, kan man klicka på det ordet och få definitionen 'mord på en dansker begået af den tyske besættelsesmagt i perioden 1943–45 som gengæld for danskernes tilsvarende attentat på et medlem af værnemagten'. Det ger en fantastisk möjlighet att lära sig ord och begrepp i ett begreppsfält.

I själva ordnätet saknas många underbegrepp (hyponymer) till *mord*, bl.a. sammansättningar med *mord* som efterled som användaren kanske hade väntat sig att finna här, t.ex. *barnemord*, *brodermord*, *folkemord* och *massemord*. Dessa hyponymer återfinns i stället i slutet av artikeln, efter idiomen. Kanske vore det bra med en upplysning om detta i användarinstruktionen.

I ordnätet över "mord" kommer det vidare upp en rubrik "andre ord med "handling" som overbegreb", alltså co-hyponymer. Under den rubriken finns en lång rad ord, t.ex. *dyrkning*, *opstart*, *luftangreb*, *bad*, *betaling* och *farvning*. Det är svårt att se vad användaren får ut av en sådan lista; problemet är självfallet att man hoppar mycket långt upp i begreppshierarkin när man går från *mord* till *handling*. Det är en tradition sedan Aristoteles att man definierar ett uppslagsord med hjälp av genus proximum, alltså det närmaste överbegreppet. I fallet *mord* kunde man då tänka sig *forbrydelse*. Om vi i stället slår upp ordet *drab*, får vi tre mer generella ord, nämligen *handling*, *forbrydelse* och *ulovlighed*. Ordnätet blir då genast mer användbart.

Att göra ett ordnät är ingen lätt sak, det finns en oändlighet av möjligheter när det gäller över-, under- och sidoordning av begrepp. Ordnätet i Den Danske Ordbog är redan en utmärkt resurs, även om det kan förbättras i flera avseenden. Den främsta svagheten är att det valda överbegreppet i många fall tycks ligga på för hög abstraktionsnivå, något som troligen ofta är ett arv från DDO:s pappersversion. I den versionen behöver det inte vara en nackdel, men det blir en nackdel när definitionerna utnyttjas för ordnätet.

Så kommer vi till den från produktionssynpunkt värdefulla rubriken "EKSEMPLER" och hittar de fasta fraserna *overlagt mord*, *et brutalt/bestialsk mord* etc. Bredvid varje fast fras finns en liten ruta med bokstaven "K"; om man klickar i den kommer man ut i korpusen och får många exempel på användningen av den aktuella frasen. Efter de fasta fraserna kommer ett fritt autentiskt exempel med källa. Man kan föra musen över den förkortning som anger

källan och se dess oförkortade namn.

Komna så långt i artikeln kan vi föreställa oss ytterligare två smärre förbättringar, en principiell och en mer tillfällig. Vi observerar först, att det inte finns någon rubrik ”GRAMMATIK”, som vid verb och många adjektiv anger att valensangivelser följer. Även många substantiv behöver egentligen valensuppgifter, men DDO är lite snål med dem (jfr Malmgren & Toporowska Gronostaj 2009). I just det här fallet kunde det vara en värdefull upplysning för många användare att det heter *mord* ***på*** *nogen*. För det andra är det kanske lite förvånande att frasen *begå mord*, som är en av de ”bästa” i listan över naboord, saknas i ordboken. (Även om frasen *begå mord* inte skulle befinna sig bland de högst placerade fraserna i listan är det värdefullt för användaren att få reda på det bästa aktiva stödverbet.)

Efter en betydelsenyans (1a) är vi framme vid ”idiomdelen”, som i det här fallet har en enda representant, *mord i øjnene*. Det åtföljs av exempel och vid de flesta idiom – dock inte detta – får man även valensuppgifter, vilket är mycket berömvärt. Alla idiom är sökbara, t.o.m. med vissa avvikelser; man kan t.ex. söka på *mord i øjet* och komma rätt.

Allra sist kommer vi till avledningar av och sammansättningar med *mord*. De är alla klickbara; från *mordaften* kommer man till den nämnda, korta artikeln **mordaften**. Om en artikel, i motsats till **mord**, har två eller flera betydelsemoment, får man genom att föra markören över sammansättningarna och avledningarna veta till vilket moment de hör.

Vad som inte minst gör utvecklingen från pappers- till nätversionen av DDO till ett kvalitativt språng är möjligheterna att navigera mellan ordboken och korpusen. Vill man ha fler exempelmeningar än ordbokens är det bara att klicka under ”Teksteksempler” i vänstermarginalen, för att få många konkordansrader med *mord*. Samma sak gäller de fasta fraserna – man klickar under ”Naboord” i vänstermarginalen och får många fler än de som står i ordboken

och kan lätt klicka sig ut i korpusen. Från korpusen kommer man lika lätt tillbaka till DDO, eller till ODS.

v

Det får väl till sist sägas att *ordnet.dk* ännu, naturligt nog, är behäftad med en eller annan barnsjukdom, som antytts här och var i det föregående. Ett ytterligare exempel, delvis från DDO-komponenten, är följande. När vi läste artikeln **skade** (substantiv) och klickade på "Teksteksempler", fick vi en mängd verbexempel (även former som bara kan vara verbformer) som vi inte fick när vi tidigare sökte efter substantivet *skade* direkt i korpusen. Det är ett fel som det bör vara ganska lätt att åtgärda.

Men ordnet.dk är redan nu en fantastisk resurs, såväl för lingvister och professionella skribenter och översättare som för den stora språkintresserade allmänheten. Vi kan bara hoppas, att vi inom en inte alltför avlägsen framtid får se lexikografiska webbsidor av samma kvalitet i Norge och Sverige.

Litteratur

Bergenholtz, Henning & Vibeke Vrang 2004: Ny dansk ordbog i seks band for sekretærer og forskere. I: *LexicoNordica* 11, 165–189.

Bergenholtz, Henning & Vibeke Vrang 2005: Den Danske Ordbog bind 2 (E–H) og 3 (I–L) – en ordbog for folket eller for akademikere? I: *LexicoNordica* 12, 169–187.

Bojsen, Else & Lars Trap-Jensen 2005: ODS, ODS-S og fremtiden. I: *10. Møde om Udforskningen af Dansk Sprog*. Århus, 58–67.

Cederholm, Yvonne & Lena Rogström 2000: Gamla ord blir som nya. OSA-databasen som källa och resurs. I: Gellerstam, M., K. Jóhannesson, B. Ralph & L. Rogström (utg.), *Nordiska studier i*

lexikografi 5. Göteborg, 66–79.
Ims, Ingunn 2008: Ordnett.no. I: *LexicoNordica 15,* 235–237.
Lundbladh, Carl-Erik 2003: Kritiken av SAOB. I: *LexicoNordica* 19, 99–117.
Malmgren, Sven-Göran 2003: Korpus 2000 – ett genombrott för tillämpad nordisk språkteknologi. I: *Språkbruk* 3/2003, 22–24.
Malmgren, Sven-Göran 2004: [Review of] Den danske ordbog, bind 1. I: *International Journal of Lexicography* 17:4, 461–467.
Malmgren, Sven-Göran & Maria Toporowska Gronostaj 2009: Valensbeskrivning i svenska ordböcker – och några andra. I: *LexicoNordica* 16, 181–196.
Pedersen, Karen Margrethe 2004: Anmeldelse af Den Danske Ordbog. I: *Danske studier* 2004, 171–183.
SO = *Svensk ordbok utgiven av Svenska Akademien.* I–II. Stockholm 2009.
Trap-Jensen, Lars 2010: One, two, many: Customization and User Profiles in Internet Dictionaries. I: Dykstra, Anne & Taneke Schoonheim (eds.): *Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010)*, 1133–1143.

(Ordböcker som nämnts huvudsakligen i egenskap av komponenter på webbsidor tas inte upp i litteraturlistan.)

Ruth Vatvedt Fjeld
professor
Institutt for lingvistiske og nordiske studier
Universitetet i Oslo
Postboks 1001 Blindern
NO-0315 Oslo
r.e.v.fjeld@iln.uio.no

Sven-Göran Malmgren
professor
Lexikaliska institutet
Institutionen för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
malmgren@svenska.gu.se

Norsk Ordbok, band VIII

Jan Terje Faarlund

Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet. Band VIII. Mugg–ramnsvart. Oslo: Det norske samlaget 2009.

Dette er ikkje ei leksikografisk melding i vanleg meining, og denne meldaren er ikkje leksikograf, men grammatikar. Oppdraget mitt var å sjå særskilt på dei grammatiske orda – eller funksjonsorda – som det skulle vera særs mange av i band VIII av *Norsk Ordbok* (NO).

Det er eit spørsmål om det bør stillast andre krav til ordboksartiklar om grammatiske ord enn til artiklar om leksikalske ord. Ettersom grammatiske ord gjerne har mindre spesifikk leksikalsk tyding, men grip inn i syntaksen på heilt andre måtar, kunne ein jo tenkja seg at ein burde leggja større vekt på syntaktiske eigenskapar enn på leksikalsk innhald ved redigeringa av desse artiklane. Eg ser ingen teikn til at det har vore gjort slike vurderingar i dei artiklane eg har sett på. Det kan i somme tilfelle ha ført til problem, som eg skal komma tilbake til.

Alle artiklane i NO startar med opplysning om ordklasse like etter sjølve oppslagsordet. Det er ikkje oppgåva til ei ordbok å diskutera ordklasseinndeling som teoretisk problem. Redaksjonen lyt bestemma seg for eit system ein gong for alle, og så halda seg konsekvent til det. Men ein grammatikar som skal melda boka, har høve til å drøfta visse prinsipielle sider ved det valde systemet. Ordklasseinndelinga i NO er den tradisjonelle, og ikkje den som Utdanningsdirektoratet har tilrådd etter innstilling frå Språkrådet, som igjen for det meste byggjer på *Norsk referansegram-*

matikk (NRG) (Faarlund/Lie/Vannebo 1997). Ein får tru at dette har å gjera med at dei fyrste banda i ordboka vart utgjevne lenge før den nye ordklasseinndelinga var aktuell. Likevel burde det ha vore mogleg å gå over til eit meir moderne system ved oppstarten av det nye prosjektet Norsk Ordbok 2014. Det ville ikkje berre ha gjort somme av dei grammatiske opplysningane i artiklane meir "tidsrette", somme ville òg ha vorte enklare. Avvika i ordklasseinndeling gjeld kategoriane **pronomen**, **konjunksjon** og **adverb**. I tradisjonell grammatikk blir dei personlege pronomena rekna som same ordklasse som determinativa (kalla "ubestemte pronomen"). NRG skil mellom pronomen og determinativ både på morfologisk og syntaktisk grunnlag; mange (ikkje alle) pronomen skil mellom ulike kasusformer, dei står normalt aleine som nominalledd, og dei har ikkje samsvarsbøying; determinativ har ikkje kasusbøying, men blir samsvarsbøygde i tal og kjønn, og dei kan stå som adledd til substantiv. Etter denne definisjonen er *nokon* eit determinativ, og ikkje pronomen. I tradisjonell grammatikk og i NO er konjunksjon brukt ikkje berre om dei eigentlege, sideordnande konjunksjonane, men også om det som no heiter **subjunksjon** (tidlegare "underordnande konjunksjon"). Dette er ikkje berre eit "nytt påfunn" av norske grammatikarar, det er i samsvar med internasjonal praksis for ordklasseinndeling. Det finst ingen grunn til at *og* og *om* (som i *om du kjem*) skal reknast til same ordklasse. Forholdet mellom adverb og preposisjon er meir problematisk og meir kontroversielt. Men mange av artiklane i dette bandet ville ha vore enklare og meir overskodelege dersom ein hadde definert dette skiljet slik det er gjort i NRG. I tradisjonell grammatikk er adverb ein slags restkategori der ein puttar alle ord som ikkje passar inn i andre kategoriar. Dermed blir alle ubøygde ord som ikkje knyter til seg adledd eller liknande, adverb. I NRG er syntaktisk funksjon hovudkriteriet ved ubøygde ord, og det finst ingen "restkategori". Dermed er *ned* preposisjon både i *Ho gjekk ned trappa* og i *Ho gjekk ned*. I begge setningane står *ned* som utfylling til *gjekk*.

Om det sjølv har ei utfylling eller ikkje, endrar ikkje på ordklasse-tilhøyret, like lite som ved verb. Ordet *opna* er verb både i *opna døra* og i *vegen opnar 1. juli*. Nett som vi har verb som berre er intransitive, som *arbeida*, har vi også preposisjonar som berre kan vera intransitive, som *nede*. Dersom ein hadde behandla *ned* (og *opp*, *over*, *på*) som preposisjonar anten dei står med utfylling eller ikkje, ville ein sleppa å skifta ordklasse straks dei får ei utfylling. Det ville også spara mykje plass, ettersom tydinga ofte er den same anten ordet har utfylling eller ei. Det er ikkje store tydingsskilnaden mellom *Han hadde hatten på* og *Han hadde hatten på hovudet.* Dette skiftet av ordklasse midt inne i ein artikkel har også ført til ein systematisk inkonsekvens ettersom det er oppgjeve ordklasse rett etter hovudoppslaget for ordet, og så ei anna ordklasse ved ein eller fleire av hovudbolkane i artikkelen: *om* er oppført som preposisjon, men uti artikkelen står det *B som adv* (kva tyder 'som' her?). Det same gjeld *ned* og *opp* adv, B prep; *nær* adj, B adv. Og det gjeld ikkje berre desse adverb/preposisjonane. Ordet *når* er såleis oppført som adverb. Under hovudbolk A står det igjen adverb, men under B har ordet plutseleg vorte konjunksjon. Eit hovudprinsipp både i ordbøker og grammatikkbøker bør vera at eit ord har same ordklasse i alle funksjonar.

Eit problem som ordboksredaktørane har meir røynsle med enn denne meldaren, og som eg veit dei er nøydde til å ta stilling til jamleg, er hierarkiseringa av dei ulike tydingane av eit ord, og like viktig i denne samanhengen, av dei ulike grammatiske funksjonane. I NO opererer dei med fire ulike nivå. Det høgste nivået er markert med romartal føre oppslagsordet. Dette skal da i prinsippet tyde at det dreier seg om ulike ord, altså homonym, som *I musling 'tufs, stakkar'* og *II musling 'blautdyr'*. Innanfor kvar artikkel finst det så tre nivå, merkte med stor bokstav, arabartal og liten bokstav, som skal skilja mellom ulike tydingsområde og bruksmåtar for kvart leksem. Eg veit at det bak desse oppstillingane ligg grundige vurderingar og drøftingar, og det er forståeleg om

redaktørane kanskje kan bli både leie og irriterte av at andre kan ha andre meiningar om inndelingane. Eg skal derfor avgrensa meg til å ta opp visse typar hierarkisering som gjeld funksjonsorda spesielt, og som ser ut til å spegla ein viss inkonsekvens i den underliggjande grammatiske analysen. Problemet kan illustrerast med handsaminga av ordet *no*. Dette har fem oppføringar. Bortsett frå *I no* med tilvising til *nu* 'vasstro', og *II no* med tilvising til *nåd* 'fred, ro', som er irrelevante homonym i denne samanhengen, er III–V ulike variantar av det same tidsordet *no*: III er substantivet, (*i neste no*), IV er adverbet, og V er interjeksjonen (*no, dei meinte det vel ikkje så gale*). Desse er ikkje homonym, men ulike bruksmåtar av same ordet. Det er eit mykje nærare samband mellom III–V innbyrdes enn mellom desse på den eine sida og I og II på den andre. Likeins er det opplagt at det er same ordet *på* vi har med å gjera i *på*: *I på* adv (*med hatten på*) og *II på* prep (*på hovudet*). Derimot er *nok* som i *ha nok pengar* og i *det var nok det beste* berre skilde med arabartal inne i artikkelen, altså på nest lågaste nivå, medan *nord* har fått to oppføringar: *I nord* m (*reisa mot nord)* og *II nord* adv (*her nord*). Det er ikkje lett å sjå konsekvensen i dette. Ein kunne jo tenkja seg at redaksjonen hadde bestemt seg for å bruka ulike romartal dersom orda høyrde til ulike ordklasser, men som vi skal sjå, er det heller ikkje gjennomført.

Artikkelen *når* er lang og interessant, og kan tena som illustrasjon på fleire av problema med artiklane om funksjonsorda. Ordet *når* er merkt adv[erb] etter hovudoppslaget. Artikkelen har to hovudbolkar, A og B. Etter A står det også adv, men etter B står det konj[unksjon]. Eg er samd i at det kan vera grunnar for å rekna dette ordet til to ulike ordklasser på syntaktisk grunnlag, adverb i *når kjem du?* og subjunksjon i *når du kjem*, men da kan ikkje den eine vera sett opp under hovudoppføringa. Merk også at dette er eit brott på det prinsippet eg nemnde ovanfor om ordklasser og hovudoppføringar. Etter oppslagsordet følgjer som vanleg ei rekkje målføreformer, men ikkje alle desse kan brukast både som ad-

verb og som subjunksjon. Forma *(h)å ner* er oppgjeven for m.a. Østre Toten, men denne kan berre vera adverb (*å ner kjæm'n?*), men ikkje subjunksjon (**å ner je ser det, så veit je det*). Målføreformer som *(h)å ner* osb. gjeld altså berre bolk A adv. Det må ha vorte noko rot i redigeringa når vi under B konj. plutseleg får fleire tilsvarande samband, *kor når* (B5d) og *når tid* (B5f), med eksempel på bruk som adverb. Og under B5d er det oppført målføreformer som *kå-ner* og *å ner*, som svara heilt til *(h)å ner* osb. i oppføringa heilt fyrst i artikkelen. I A2a står det at *når* tyder 'alltid, støtt' i sambandet *når som helst*; det er vel heile sambandet som tyder 'alltid, støtt'. Elles skal redaktøren av artikkelen ha ros for at ho under eit eige punkt (B1a) seier klart at *når* blir brukt 'om fortid som gjeld eitt høve', altså der skulegrammatikken og sjølvoppnemnde språkguruar (ein treng ikkje nemna namn!) insisterer på at det lyt heita *da*. Det har sjølvsagt ikkje grunnlag i norsk talemål, og bør heller ikkje gjelda nynorsk standardmål.

Ei fullstendig ordbok skal også ha med syntaktiske opplysningar, og det finst det mykje av her. Ordklasseopplysningane gjeld sjølvsagt syntaksen, og det skiljet mellom adverb og preposisjon som eg kritiserte ovanfor, gjev jo også syntaktiske opplysningar om ordet. (Dersom ein skulle bruka den nye ordklasseinndelinga, måtte da slike opplysningar gjevast på andre måtar.) Men i dette bandet finst det eit par interessante verb, der viktige syntaktiske opplysningar manglar. Det gjeld verba *pla* og *pleia*. Dei tyder om lag det same, og begge tek infinitiv som utfylling. Dette er ikkje nemnt, men det går fram av døma. Det er likevel ein interessant skilnad mellom dei to verba: *pla* er alltid følgt av infinitiv utan *å*, slik det går fram av alle døma i artikkelen. Dette blir da syntaktisk å rekna som eit (modalt) hjelpeverb. I den stutte artikkelen om *pleia* er det berre to døme, eitt med *å* og eitt utan. Med så få døme er det sjølvsagt ikkje mogleg å avgjera om variasjonen er geografisk betinga. Men det er interessant at setninga utan *å* er nekta, dette ser ut til å vera eit mønster og eit interessant forhold ved

dette ordet (Johannessen 2002). Eit anna ord på *p* med ein pussig syntaks er *pluss*. Som NO opplyser, blir det brukt for å addera tal. Men det blir også brukt for å "addera" setningar, og det pussige er da at den andre setninga må vera ei at-setning, jamvel om den fyrste ikkje er det: *Maten var for dyr, pluss at sørvisen var elendig* (Lie 1979, Julien 2009).

Semantikk er det sjølvsagt rikeleg med her. Og naturleg nok gjeld det leksikalsk semantikk, ordinnhald. Men visse typar ord har også semantiske eigenskapar som går ut over ordet sjølv. Det gjeld mellom anna ord for mengd og ord for modalitet, særleg når dei opptrer saman med nekting. Eg vil spesielt trekkja fram to slike ord frå dette bandet. Det eine er det modale hjelpeverbet *måtta*. Dette er eit mangfelt verb, som det har vore skrive avhandlingar om (m.a. Engh 1975), og det er ikkje rom for alle nyansar i ein ordboksartikkel. Men ein kunne venta at denne artikkelen var litt klarare strukturert ut frå to dimensjonar: For det fyrste lyt ein skilja mellom den deontiske og den epistemiske bruken (eg seier ikkje at redaktøren skal bruka desse tekniske termane, men dei kunne ha vore til hjelp for henne i redigeringa av artikkelen). Artikkelen har sju delar nummererte med arabartal, dei fleste med underpunkt (ingen hovudbolkar med stor bokstav). Den epistemiske bruken er omtala og eksemplifisert i 6a, eitt typisk døme er *vi må ha teki feil*. Resten av punkt 6 og alle dei andre punkta handlar om den deontiske bruken, som i *alle må levera sjølvmelding* eller *må eg få det?* Dei to siste setningane representerer også to ulike tydingsområde, 'vera nøydd til, ljota' og 'få lov til'. Det er nokre døme på den siste bruken under punkt 4a, saman med andre bruksmåtar. Verbet *måtta* i denne tydinga er vanleg i dansk og i danskprega norsk. I mange norske dialektar blir vel *må* brukt slik mest (eller berre) i nekta setningar: *du må ikkje drikka for mykje i kveld*. Ved nekting får vi derfor ei tvityding som vi ikkje har elles: *du må ikkje skriva under* kan tyda anten *du lyt ikkje skriva under (du kan la vera om du vil)* eller *du må la vera å skriva under*. Sambandet av eit

mengdeord som *nokon* og nekting har også interessante semantiske konsekvensar. I ei så pass omfattande ordbok som NO kunne det vera plass for å visa skilnaden mellom *ikkje nokon kunne svara på det* og *nokon kunne ikkje svara på det.*

Det ser ut til å vera ei lov om ordbøker at det er eit omvendt proporsjonalt tilhøve mellom lengda på ordet og lengda på ordboksartikkelen. Dermed er det også ein lang artikkel om *og*. Det er imponerande kor mange ulike bruksmåtar redaktøren har funne fram til, jamvel om det i somme tilfelle snarare er tale om ulike typar innhald i dei to konjunkta enn ved ordet *og* i seg sjølv. I denne artikkelen ser vi også noko av problemet med ei redigering på ikkje-grammatisk grunnlag: ein så pass viktig og syntaktisk interessant bruk av *og* som den vi finn ved pseudokoordinering (ikkje ein term for NO), som i *sitja og eta*, er gøymt bort saman med andre konstruksjonstypar i eit underpunkt på lågaste nivå, 2o. I tråd med den deskriptive, ikkje-normative linja, er også *og* som infinitivsmerke teke med, det er jo utan tvil ein vanleg bruk av ordet i dagens norsk. Det einaste eg sakna i denne artikkelen, er *og* som uttrykk for samanhengen mellom to omgrep, altså snittet i staden for unionen, for å seia det i logikkspråk. Dersom eg eit semester skulle tillysa eit seminar om "Språk og kjønn", ville ikkje studentane venta seg eit halvt semester om språk og så eit halvt semester om kjønn, men eit heilt semester om kva dei to har med kvarandre å gjera. Ein detalj til slutt: Eg undrar meg litt over at adverbet *òg* er skrive utan aksent som oppslagsord (i innleiinga til bandet skriv prosjektdirektøren *òg* med aksent i samsvar med offisiell rettskriving).

Jamvel om det har vore mange kritiske merknader til artiklane om dei grammatiske orda, har eg ikkje lyst til at lesarane av denne meldinga skal sitja att med inntrykk av at redaktørane av NO ikkje har greie på grammatikk. Artiklane inneheld ei mengd gode og riktige tolkingar, analysar og forklaringar. For det aller meste er artiklane godt strukturerte, og dei ulike bruksmåtane er rikt il-

lustrerte. Her har eg berre gripe fatt i dei få tilfella der eg meiner det med fordel kunna ha vore gjort annleis – og det er eit fåtal av tilfella. Som ved tidlegare utgjevnader kan ein berre lata seg imponera over omfang, kvalitet, og ikkje minst presisjon i arbeidet åt det store ordboksprosjektet.

Litteratur

Engh, Jan 1975: Må *og* kan *med objektiv lesning*. Upubl.

Johannessen, Janne Bondi 2002: Negative Polarity Verbs in Norwegian. I: *Working Papers in Scandinavian Syntax*, 33–74.

Julien, Marit 2009: Plus(s) at(t) i skandinaviska – en minimal matris. I: *Språk och stil* 19, 124–141.

Lie, Svein 1979: PLUSS. I: *Norskrift* 24, 50–60.

NRG = Jan Terje Faarlund/Svein Lie/Kjell Ivar Vannebo 1997: *Norsk referansegrammatikk*. Oslo: Universitetsforlaget.

Jan Terje Faarlund
professor
CSMN, Universitetet i Oslo
Postboks 1020 Blindern
NO-0315 Oslo
Noreg

”Ordaboken moste tryckias”

Anna Helga Hannesdóttir

Jesper Swedberg: *Swensk Ordabok. Utgiven efter Uppsala-handskriften, med tillägg och rättelser ur övriga handskrifter, av Lars Holm.* Skara stiftshistoriska sällskaps skriftserie nr 46, och Acta Bibliothecae Scarensis. Skrifter utgivna av Stifts- och landsbiblioteket i Skara, nr 12. Skara: Stiftelsen för utgivande av Skaramissalet 2009, 719 s.

Nu har Lars Holm uppfyllt biskop Jesper Swedbergs (1653–1735) sista vilja och gett ut den svenska ordbok som han arbetade med under en stor del av sitt liv. Trots att verket tills nu har legat opublicerat har dagens språkhistoriker, genom Holms uthålliga umgänge med och många presentationer av det, haft goda möjligheter att stifta bekantskap med olika aspekter av denna 1700-talsordbok. I sin doktorsavhandling (Holm 1986a) och ett tiotal bidrag, bland annat i denna tidskrift och i volymerna från konferenserna om lexikografi i Norden, har Holm redovisat sina studier i Swedbergs ordboksarbete. Utgåvan innehåller också en bibliografi: ”Lars Holm om Jesper Swedbergs *Swensk Ordabok*” (Holm 2009:44 f.). Han har också sammanfattat sin Swedbergsforskning i populärvetenskaplig form (Holm 2008). Då själva ordboken således är väl utforskad ut lexikografiskt perspektiv avstår jag här ifrån att fördjupa mig i enskilda lexikografiska aspekter av själva ordboken. Istället är det utgåvan i sin helhet som fokuseras.

1. Utgåvans innehåll

Utgåvan består av två delar. Första delen utgörs av Lars Holms

introduktion (85 sidor). Där presenterar han kort Jesper Swedbergs liv och gärning, beskriver sina redigeringsprinciper och redovisar i avkodad form de källor till ordförrådet som Swedberg anger i ordboksartiklarna. Vidare översätter och förklarar Holm det latinska metaspråket i ordboken. Introduktionen avslutas med exempelsidor i faksimil av de fem bevarade handskrifterna av ordboken.

Den andra delen innehåller editionen: ”Jesper Swedberg: Swensk ordabok. R 589 e, Uppsala universitetsbibliotek”. Editionen omfattar hela handskriften: från titelsida och kringtexter till Swedbergs förord (16 sidor) och själva ordbokstexten (589 sidor). Holms syfte med utgåvan är att ”göra Ordaboken tillgänglig för läsare och brukare som vill fördjupa sig i den svenska som talades och skrevs för 300 år sedan” (Holm 2009:26).

2. Holms introduktion

Introduktionen inleds med en kort översikt över den tidiga svenska lexikografin och dess förankring i den kontinentala, latinska lexikografitraditionen. Holm konstaterar att två skilda traditioner kan urskiljas, dels den som har kallats ”den patriotiska traditionen”, dels ”den nyttoinriktade traditionen” (Ralph 2000:14 f.). Dessa två angreppssätt på det svenska ordförrådet menar Holm företräds vid 1700-talets början av var sin biskop: Haqvin Spegel och Jesper Swedberg.

2.1. Holm om Swedberg

Holm ger en kortfattad men initierad – och bitvis ömsint – beskrivning av Swedbergs liv och hans karriär inom de fyra ”institutioner” som hans verksamhet kom att präglas av och kretsa kring: skolan, det teologiska studiet, den nationella scenen där han verkade som hovpredikant och professor och slutligen den svenska

kyrkan. Swedbergs insatser i den bibelrevision som resulterade i Karl XII:s bibel beskrivs kortfattat, liksom hans inblandning i arbetet med att ta fram en svensk rikspsalmbok, en inblandning som inte blev särskilt ärorik för den då blivande biskopen. Den swedbergska psalmboken trycktes 1694 men hela upplagan drogs omedelbart tillbaka – främst av teologiska skäl (se Alling 2008). Också hans övriga språkvetenskapliga produktion omnämns: *Schibboleth. Swenska språketz rycht och richtighet* (1716), vars stavningsregler eldade under den ”giganternas kamp” om svenskans stavning som han sedan psalmboksaffären utkämpade med Urban Hiärne (Ohlsson 1992:11), och hans betydligt kortare grammatik, *En kortt Swensk Grammatica* (1722). Det var också Swedbergs ortografiska ställningstaganden som till slut tycks ha omöjliggjort utgivningen av ordboken.

I introduktionen berörs också Swedbergs brevväxling med bokcensorn Upmarck-Rosenadler och Kanslikollegiets handläggning av frågan om stöd till utgivning av ordboken. Utöver den kontroversiella ortografin, som redan tidigare använts som argument mot utgivande, anförde Kanslikollegiet ekonomiska argument, och när Swedberg i början av 1730-talet efter att ha aktualiserat den indragna psalmboken tillrättavisades av drottning Ulrika Eleonora, var utgivningen av ordboken även politiskt omöjlig.

I tidigare sammanhang har Holm presenterat ordbokens tillkomsthistoria betydligt mera detaljerat än vad han gör i utgåvan (Holm 1984 och 2008) och det är naturligtvis inte rimligt att där återge alltför mycket av det som finns publicerat. Men för den läsare som inte i förväg är förtrogen med ordboken och med Swedbergs språkliga insatser är introduktionen lite väl knapphändig för att riktigt göra biskopens arbete rättvisa.

2.2. Holm om manuskripten

Holm redogör för de fem manuskript som finns av ordboken, deras

tillkomst, nuvarande omfång och skick samt antagna inbördes relationer. En noggrannare genomgång återfinns i Holm (1985). Till grund för utgåvan läggs den avskrift som sonen Jesper Swedenborg färdigställde och som nu förvaras i Uppsala universitetsbibliotek. Denna kom Swedberg själv att ha som arbetsexemplar, där han under resten av sitt liv införde tillägg och korrigeringar.

Också andra har kompletterat Jesper den yngres avskrift. Bland dessa finns antagligen Petrus Schenberg. Han utgår ifrån manuskriptet till Swedbergs ordbok i arbetet med den svensklatinska delen av sin stora *Lexicon latino-svecanum* 1739 (Schenberg 1739:[2]). Hur han fick tillgång till manuskriptet har Holm tidigare diskuterat (Holm 1984:118 f.). Vad som däremot undgått forskningen är att den inställning till modersmålet som framgår av Schenbergs syfte med ordboken, nämligen att den ska ”fungera som stöd vid gymnasie- och skolungdomens studium av latinet och modersmålet” inte är hans (Hannesdóttir 1998:196). Denna pedagogiska syn kan i själva verket tillskrivas Swedberg, som hyste den för den tiden brådmogna ”önskan om att låta ’alt lärande ske på modersmåhlet’” (Holm 2009:22). Det skulle ännu dröja länge tills undervisning i modersmålet blev en uppgift för skolan.

2.3. Holm om ordboken

Efter en överskådlig presentation av ordbokstextens informationskategorier, följer Holms betraktelse över hur några av Swedbergs språkliga käpphästar kommer till uttryck i ordboken. Ortografin är en sådan. Inledningsvis i det ”Prof på en Swensk Ordabok” som trycktes som kapitel 18 i *Schibboleth* (1716) konstaterar Swedberg att ”Wåre Swenske ord skrifwas åtskilligt af åtskilliga”. Redan underrubriken till kapitlet innehåller hans programförklaring: ”at orden på ett wißt sett altid skrifwas kunna och böra” (Swedberg 1716:421). Hur han för egen del uppfyllde denna vision exemplifierar Holm bl.a. genom att diskutera hur han tillämpar en enhetlig ortografisk markering av lång vokal.

Bland andra språkliga drag som Swedberg ägnat särskild uppmärksamhet i ordboken och som Holm förtjänstfullt lyfter fram är distinktionen mellan de gamla svaga maskulina och feminina substantivens grundform och oblika form. Dessa uppslagsord återges i nominativ och också den oblika formen ges i artikeln – om denna är belagd i bibeln. Som Holm också visar gör Swedberg en pionjärinsats genom att ta med ett stort antal löst sammansatta partikelverb, och läsaren uppmärksammas på den omsorgsfulla redovisningen av såväl betydelsevariation som uttrycksvariation. Hur Swedberg hanterade polysemi har Holm behandlat närmare i annat sammanhang (Holm 2005a), liksom uttrycksvariation eller möjlig synonymi (Holm 2005c).

Holm ägnar ordbokens källor och ordskatt ett eget avsnitt, och också på detta område har han tidigare genomfört specialstudier. Bibelns ordförråd är väl täckt i ordboken (se också Holm 2005b); inemot hälften av uppslagsorden har Swedberg hämtat där. Dessa ord beläggs i de flesta fall med hänvisningar till ett eller flera angivna bibelställen. Dessutom har samtida svenska författares skrifter inom vitt skilda områden skattats på ord som – med angivande av källan – upptagits i ordboken. Swedberg förtecknar också ord av det slag som han antagligen inte har belagt i skriftliga källor eller som han i varje fall inte redovisar källorna till, bl.a. ord som ingår i det vardagliga, familjära och även låga ordförrådet. Förutom bibelorden har Holms tidigare studier av ordbokens ordförråd ägnats åt biskopens uppsättning av fula ord (Holm 1986) och främmande ord (Holm 1988).

Avslutningsvis i sin presentation av Swedbergs ordbok argumenterar Holm för att denna gamla ordbok har en del att erbjuda dagens läsare – även andra än språkhistoriker med olika specialintressen. Ordboken förtecknar inte bara det tidiga 1700-talets allmänna ordförråd utan också den då aktuella terminologin inom vitt skilda områden. Där märks rättskipning, krigsväsende och kyrka liksom mer basala områden som tidens matvanor och

klädedräkt. Även benämningar på svenskarnas egna och deras husdjurs krämpor och sjukdomar redovisas. Genom utgåvan har ordbokens juridiska ordförråd nu kunnat inlemmas i en undersökning av den juridiska terminologins etablering (Rogström u.a.).

2.4. Holm om sin utgåva

Holm väljer vad han kallar ”brukbarhetsprincipen” för sin utgåva. Det viktiga för honom är inte att redovisa en enskild skrivares (eller avskrivares) egenheter. Han vill i stället ”restaurera, återskapa en Swensk Ordabok som så nära som möjligt överensstämmer med den ordbok som Jesper Swedberg ville ge ut” (Holm 2009:37). Swedbergs ordbok ska vara så tillgänglig som möjligt för nutida läsare. Det är innehållet i Uppsalahandskriften som ges ut, liksom i denna ordnat i två spalter på sidan, men artiklarnas form, struktur och notation följer i stor utsträckning Swedbergs eget manuskript.

I vissa avseenden normaliserar Holm texten: Diplomatarisk trohet får stå tillbaka för hans strävan efter konsekvens och tydlighet. Han inför därför en enhetlig grafisk form för uppslagsorden. Istället för att återge ”den regellösa växlingen” i handskriften mellan stor och liten begynnelsebokstav i uppslagsorden ändrar han förutom i egennamn genomgående versal till gemen (Holm 2009:39). De särskrivningar och sammansättningar med divis som förekommer sporadiskt och slumpmässigt i manuskriptet skriver han ihop. Också interpunktionen i artiklarna normaliseras liksom olika allografer för vissa grafem (exempelvis några av allograferna för <s>). Andra behålls konsekvent. Denna i princip grafiska normalisering är i sammanhanget motiverad och väl avvägd.

Också ifråga om textens disposition på handskrifts- respektive trycksidorna tillämpar Holm en pragmatisk princip. Handskriftens paginering och satsyta för övrigt bryts upp och anpassas till utgåvans format. Uppsalahandskriftens makrostruktur behålls –

också de avvikelser från den alfabetiska ordningen som vissa av Swedbergs egna ortografiska ändringar gett upphov till.

I utgåvans notapparat redovisas de tillägg i texten som gjorts i förhållande till R 589 e. Även variantformer i de andra handskrifterna förtecknas i noterna och där framgår också varifrån editionens variant är hämtad om källan är en annan än just R 589 e. Likaså redovisas i noterna Holms egna rättelser och förekomsten av andra händer än Swedbergs och sonen Swedenborgs i manuskripten. Holm förklarar sitt notsystem och exemplifierar med noter hämtade ur utgåvan och han kommenterar på ett tydligt och pedagogiskt sätt de olika typer av information som ges. Noterna är, på sedvanligt sätt, enkelriktade i och med att det i artiklarna inte finns någon markering av att ytterligare information finns att hämta i en not och noterna återger inte uppslagsordet om inte informationen gäller detta specifikt. Den läsare som är intresserad av eventuella variantformer är därför hänvisad till att via radnumret söka i ordbokstexten. Fördelarna med denna lösning är främst grafiska och estetiska; variantapparaten stör inte den lexikografiska informationen. En nackdel med de enkelriktade noterna är dock att det krävs mer av den läsare som är ute efter att skaffa sig en uppfattning om exempelvis de olika handskrifternas inbördes relationer i enskilda fall.

Holm har verkligen lyckats i sin strävan att åstadkomma en för den nutida läsaren överskådlig och lättillgänglig edition. Hans varsamma normalisering av den fria variation i ortografi och notation, som den ännu inte normaliserade svenskan lämnade utrymme för i 1700-talets början, bidrar stort till utgåvans tilldragande yttre. Bilderna 1 och 2 nedan visar början på bokstavskapitlet *F* i Swedbergs respektive Holms version.

2.5. Holm om Swedbergs källor

I ordboksartiklarna hänvisar Swedberg, liksom den tidens lexiko-

grafer, i starkt förkortad form till de verk han har hämtat sina uppslagsord eller ekvivalenter i. Holm inte bara redovisar Swedbergs förkortningar – han har lyckats med bedriften att identifiera de flesta av de omkring 300 verk som åberopas. För dessa titlar ges kortfattade bibliografiska uppgifter, ibland också en kort kommentar om verket ifråga. Läsaren får också reda på vilka ord som belagts i respektive verk, utom när det gäller bibeln och psalmboken. Ordförrådet i dessa utgör ju själva utgångspunkten för Swedbergs ordbok. Holm ransonerar också orduppgifterna till högst tio ord från Swedbergs andra fyra huvudkällor: den egna *Schibboleth*, Spegels *Glossarium*, kyrkoordningen från 1571 och Wankijffs bibel (1674 och 1688).

Holm gör en distinktion mellan Swedbergs källor (för uppslagsorden) och referenser (för ekvivalenterna eller annan relevant information), men denna indelning genomförs inte explicit i förteckningen. I de flesta fall är det entydigt vad som har hämtats i det anvisade verket, såsom i fallet ”hirtar sig” (Holm 2009:335) som inte förses med annan information än en hänvisning till ”Stiernh[ielms] Glossar” och ”gudz werck och h.”. Att sedan det sistnämnda verket inte återfinns under *gudz* i den alfabetiskt ordnade förteckningen utan ordnat på sin upphovsman: ”(Spegels) guds werck och h. / hwila” (Holm 2009:73) är ren lapsus. För den som vill studera de åberopade arbetena närmare, redovisar Holm också vid vilka bibliotek en del av de i dagens Sverige mindre kända verken har påträffats eller – vilket kan vara lika viktigt – sökts förgäves.

De ytterligare förteckningar Holm bistår den moderne läsaren med är dels en lista över de förkortningar Swedberg använder för Bibelns böcker, dels en komplett latinsk-svensk ordlista över det latinska metaspråket. Ordlistan är generös på så sätt att här redovisas inte en abstraherad grundform av de latinska ord som förekommer utan de former som faktiskt är belagda, ofta med uppgift om vad det är för form. För *habent* anges ”har (3 pers. plur.)”

och för *habet* ”har (3 pers. sing.)” (Holm 2009:83). Biskopen hade säkert förfärats men till skillnad från de användare som ordboken ursprungligen var avsedd för kan många av de läsare, som han med Holms hjälp till slut får hålla tillgodo med, inget latin alls.

3. Editionen av Jesper Swedbergs svenska ordabok

När nu biskopens ordbok äntligen tryckts, är det trots allt inte den ordbok som förelåg i färdigt skick runt 1720. Istället korrigerades de fel och brister som Swedberg (och andra) upptäckte, och ordförrådet utökades såväl med fula ord som med andra ord som inte finns i Swedbergs eget manuskript.

Som framgått ovan är det Uppsalahandskriften (R 589 e) som Holm har utgett i sin helhet. Efter titelsidan inleds den med en anteckning av Swedberg och ett latinskt citat ur Plinius den äldres *Historiæ Naturalis*. Därefter följer ett utlåtande om ordboken, undertecknat J. Rosenadler. Huruvida utlåtandet verkligen har formulerats av Rosenadler har Holm tidigare diskuterat (Holm 1984:112 f. och 1985:8 f.). Utlåtandet följs av en sida med ordbokens hela titel och en lista över de förkortningar som används för Bibelns böcker. I utgåvan infogas handskriftens titelsida och sidan med de upplösta förkortningarna i faksimil. Efter Swedbergs förord och själva ordbokstexten avslutas R 589 e och utgåvan med en ”Förtekning på the skrifftennes språk och rum, som af anförda wackra lärare äro förklarade!”.

Holms textkritiska utgåva är snygg. Formatet är det som ordboken antagligen hade fått om den hade getts ut när den var färdig. Stilen är tydlig och artiklarnas mikrostruktur är överskådlig genom de olika stilsorterna. För den som vill följa textens fördelning på manuskriptets sidor anges manuskriptets sidnumrering och sidbrytning i marginalen.

Vi har nu fått tillgång till Swedbergs ordförråd och – genom Holms försorg – uppgifter om var orden förekommer. I många fall får vi också exempel på konstruktioner de ingår i. Som Holm påpekar ställs emellertid den nutida läsaren inför en stor svårighet: Vad exakt betyder egentligen dessa ord? De svenska uppslagsord som förklaras förses alltid med en ekvivalent eller en parafras på latin. De förklaras sällan på svenska annat än när homonymer disambigueras och indirekt när läsaren hänvisas till en synonym. På så sätt är Swedbergs ordbok en tydlig manifestation av latin-

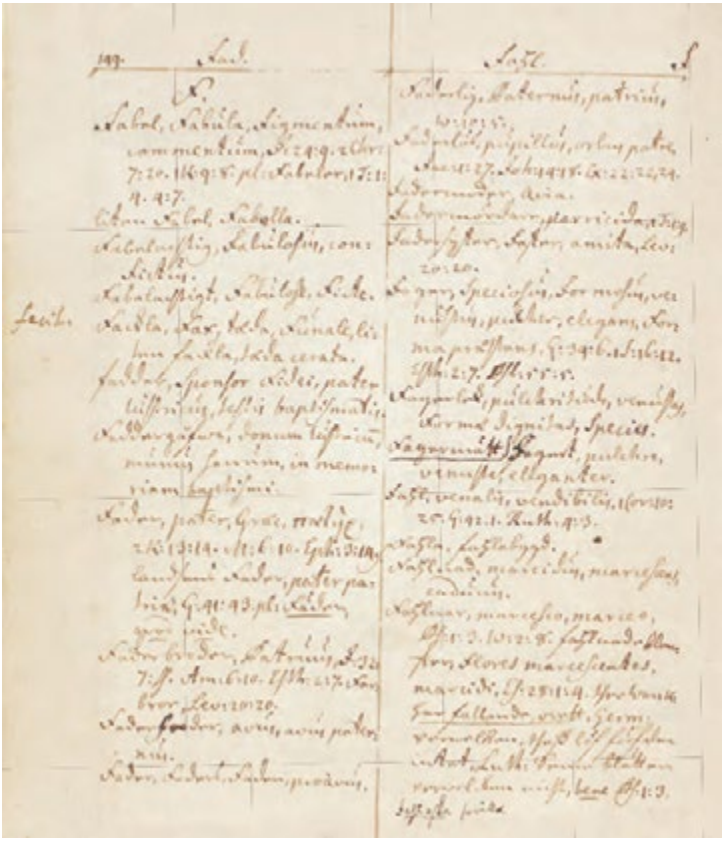
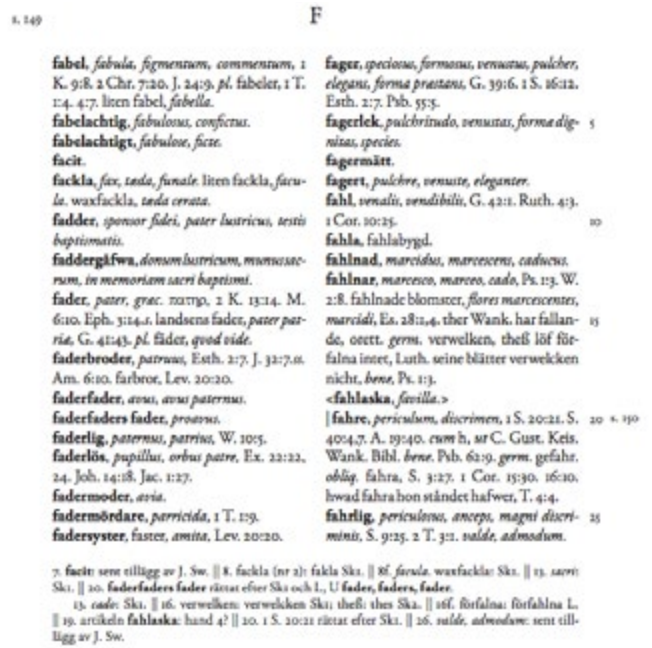


Bild 1: Början på bokstaven *F* i R 589 e

ets starka ställning i 1700-talets Sverige. För att komma åt den betydelse som ett ord hade för 300 år sedan behöver man, påpekar Holm, gå via latinet och återöversätta de latinska ekvivalenterna och parafraserna till dagens svenska. Därigenom menar han skulle också ”språkhistoriskt falska vänner” kunna avslöjas (Holm 2009:43).



s. 149 F

fabel, *fabula, figmentum, commentum,* 1 K. 9:8. 2 Chr. 7:20. J. 24:9. *pl.* fabelet, 1 T. 1:4. 4:7. liten fabel, *fabella.*
fabelachtig, *fabulosus, conficius.*
fabelachtigt, *fabulose, ficte.*
facit.
fackla, *fax, teda, funale.* liten fackla, *facula.* waxfackla, *teda cerata.*
fadder, *sponsor fidei, pater lustricus, testis baptismatis.*
faddergåfwa, *donum lustricum, munuscu-rum, in memoriam sacri baptismi.*
fader, *pater, græc.* πατηρ. 2 K. 13:14. M. 6:10. Eph. 3:14,1. landsens fader, *pater patriæ,* G. 41:43. *pl.* fäder, *quod vide.*
faderbroder, *patruus,* Esth. 2:7. J. 32:7,11. Am. 6:10. farbror, Lev. 20:20.
faderfader, *avus, avus paternus.*
faderfaders fader, *proavus.*
faderlig, *paternus, patrius,* W. 10:5.
faderlös, *pupillus, orbus patre,* Ex. 22:22, 24. Joh. 14:18. Jac. 1:27.
fadermoder, *avia.*
fadermördare, *parricida,* 1 T. 1:9.
fadersyster, faster, *amita,* Lev. 20:20.

fager, *speciosus, formosus, venustus, pulcher, elegans, forma præstans,* G. 39:6. 1 S. 16:12. Esth. 2:7. Psb. 55:3.
fagerlek, *pulchritudo, venustas, forma dignitas, species.*
fagermätt.
fagert, *pulchre, venuste, eleganter.*
fahl, *venalis, venditbilis,* G. 42:1. Ruth. 4:3. 1 Cor. 10:25.
fahla, fahlabygd.
fahlnad, *marcidus, marcescens, caducus.*
fahlnar, *marcesco, marceo, cado,* Ps. 1:3. W. 2:8. fahlnade blomster, *flores marcescentes, marcidi,* Es. 28:1,4. ther Wank. har fallande, oeett. *germ.* verwelken, theß löf förfalna intet, Luth. seine blätter verwelcken nicht, *bene,* Ps. 1:3.
<fahlaska, *favilla.*>
|**fahre**, *periculum, discrimen,* 1 S. 20:21. S. 40:4,7. A. 19:40. *cum* h, *ut* C. Gust. Kels. Wank. Bibl. *bene.* Psb. 62:19. *germ.* gefahr. *oblig.* fahra, S. 3:27. 1 Cor. 15:30. 16:10. hwad fahra hon ståndet hafwer, T. 4:4.
fahrlig, *periculosus, anceps, magni discriminis,* S. 9:25. 2 T. 3:1. *valde, admodum.*

7. **facit**: sere tillägg av J. Sw. || 8. fackla (nr 2): fakla Sk1. || 8f. *facula.* waxfackla: Sk1. || 13. *cado*: Sk1. || 10. **faderfaders fader** rättat efter Sk1 och L, U **fader, faders, fader**.
13. *cado*: Sk1. || 16. verwelken: verwelcken Sk1; theß: ther Sk2. || 16f. förfalna: förfahlna L. || 19. artikeln **fahlaska**: hand 4? || 20. 1 S. 20:21 rättat efter Sk1. || 26. *valde, admodum*: sent tillägg av J. Sw.

Bild 2: Början på bokstaven *F* i Holms utgåva av *Swensk Ordabok*

Ytterligare ett användningsområde för ordboken som Holm föreslår vore att parallelläsa den med SAOB för att korrigera den sistnämndas uppgifter om förstabelägg.

4. Holms *Swensk ordabok*

Det är naturligtvis omöjligt att avgöra vilken inverkan ordboken faktiskt hade haft på den svenska lexikografins utveckling om den hade tryckts då den omkring 1720 förelåg färdig. Att den nu, närmare tre sekler efter sin tillkomst, finns utgiven kommer heller inte att påverka den framtida svenska lexikografin. Däremot är utgåvan allmänt språkhistoriskt, specifikt lexikografihistoriskt och filologiskt intressant. Förutom den tryckta boken finns den också åtkomlig i digitaliserad form i Litteraturbanken (se litteraturförteckningen).

När det gäller den *Swensk ordabok* som nu äntligen lämnat trycket finns det en asymmetri mellan Holms – i vissa avseenden – summariska introduktion och hans grundliga utgåva. Den exposé över Swedbergs språkvetenskapliga gärning som ges i introduktionen hade tjänat på att problematiseras och relateras till dagens forskningsläge. När Holm började sitt umgänge med biskop Swedbergs ordbok var den svenska lexikografins utveckling ett outforskat fält. Sedan dess har det språkliga och specifikt lexikografiska klimat som Swedberg verkade i undersökts ur olika perspektiv. I Stig Örjan Ohlssons forskning kring Urban Hiärnes språkvetenskapliga skrifter figurerar naturligt nog också Swedberg. Grammatikbegreppets utveckling har utforskats och då har också Swedbergs grammatik (1722) satts in i detta sammanhang (Haapamäki 2002). Sedan Holm disputerade på Swedbergs ordbok har de stora dragen i den svenska lexikografins framväxt i princip kartlagts. Därvid har den tidiga latintraditionens utveckling och arvet efter *Lexicon Lincopense*, som ju Swedberg har att förhålla sig till, redovisats (Johansson 1997), och de gamla ordböckernas värde som språkhistoriska källor har problematiserats (Ralph 1992). Dessa perspektiv hade Holm genom hänvisningar kunnat leda läsaren till.

Här har endast den tryckta utgåvan av ordboken beaktats. Som framgått ovan finns ordboken också tillgänglig som pdf-fil i Litteraturbanken. Formatet tillåter en rad sökningar och inbjuder på så vis till studier som är tidsödande att genomföra med boken som underlag. Det är exempelvis tacksamt att använda filen för ordförrådsstudier av olika slag. En blixtstudie av kvinnobeteckningar med initialt *b* och bildade med avledningssuffixet *-erska* gav direkt sex träffar. Ett av orden, *bislåperska*, förekommer också i biskopens förord. I ordboken får det ekvivalenterna *concubina* och *pellex*. Genom en sökning på *concubina* påträffas synonymerna *frilla* och *sengakamerat*, som entydigt bekräftar det semantiska fält *bislåperska* hör hemma i. Även de rent filologiska aspekterna blir lättåtkomliga i pdf-filen genom sökningar på exempelvis ”sena tillägg”, ”rättat efter” m.m.

I sin slutgiltiga form är *Swensk Ordabok* lika mycket Holms verk som Swedbergs. Swedberg tillhandahöll ett material som Holm nog har ägnat lika lång tid åt sammanlagt som upphovsmannen. Resultatet är på det hela taget imponerande. Utgåvan är noggrann och konsekvent samtidigt som den är lättillgänglig och elegant. I sin fysiska form pryder boken sin plats i bokhyllan och i den digitala formen inbjuder den till fortsatta studier i det tidiga 1700-talets svenska. Swedberg hade definitivt inte gjort det bättre själv.

Litteratur

Alling, Gunnar 2008: Swedbergs försvar. En studie av Jesper Swedbergs levernesbeskrivning med särskild hänsyn till psalmbokstriden 1694. I: *Jesper Swedberg – en antologi*. Red. Johnny Hagberg. (Skara stiftshistoriska sällskaps skriftserie 35.) Skara, 301–337.

Haapamäki, Saara 2002: *Studier i svensk grammatikhistoria*. Åbo: Åbo Akademis förlag.

Hannesdóttir, Anna Helga 1998: *Lexikografihistorisk spegel. Den enspråkiga svenska lexikografins utveckling ur den tvåspråkiga*. (Meijerbergs arkiv för svensk ordforskning 23.) Göteborg.

Holm, Lars 1984: Swedbergs Swensk Ordabok – yttre dokumentation. I: Holm 1986a.

Holm, Lars 1985: Swedbergs Swensk Ordabok – handskrifterna. I: Holm 1986a.

Holm, Lars 1986a: *Jesper Swedbergs Swensk Ordabok – bakgrund och tillkomsthistoria*. Uppsala.

Holm, Lars 1986b: När biskopen blev gammal blev han ful i mun eller: De sena tilläggen i Swedbergs Swensk Ordabok (Uppsala-handskriften): I: *Svenska i tid och otid*. Vänskrift till Gun Widmark från doktoranderna i Uppsala. Uppsala.

Holm, Lars 1988: Swedberg och de främmande orden. I: *Studier i svensk språkhistoria*. Red. Gertrud Pettersson. (Lundastudier i nordisk språkvetenskap. A 41.) Lund, 104–118.

Holm, Lars 2005a: Polysemi i Swedbergs Swensk Ordabok. I: *Nordiske studier i leksikografi* 7. (Rapport frå konferanse om leksikografi i Norden. Volda 20.–24. mai 2003, red. av Ruth Vatvedt Fjeld og Dagfinn Worren.) Oslo, 186–203.

Holm, Lars 2005b: Swedbergs Swensk Ordabok och Bibeln. I: *Den gamla översättningen. Karl XII:s bibel och dess receptionshistoria.* Föredrag vid en konferens i Lund den 21–25 februari 2003 anordnad av Kungl. Humanistiska Vetenskapssamfundet i Lund. Redigerad av Tord Larsson och Birger Olsson. Lund.

Holm, Lars 2005c: Uttrycksvariation i Swedbergs Swensk Ordabok. I: *Studier i svensk språkhistoria* 8. Red. Cecilia Falk och Lars-Olof Delsing. (Lundastudier i nordisk språkvetenskap. A 63.) Lund, 125–135.

Holm, Lars 2008: *Jesper Swedbergs* Swensk Ordabok. I: *Jesper Swedberg – en antologi*. Red. Johnny Hagberg. (Skara stiftshistoriska sällskaps skriftserie 35.) Skara, 339–406.

Holm, Lars 2009: *Jesper Swedberg:* Swensk Ordabok*. Utgiven efter Uppsala-handskriften, med tillägg och rättelser ur övriga handskrifter, av Lars Holm.* (Skara stiftshistoriska sällskaps skriftserie nr 46, och Acta Bibliothecae Scarensis. Skrifter utgivna av Stifts- och landsbiblioteket i Skara, nr 12.) Skara.
Även: <http://litteraturbanken.se/#forfattare/SwedbergJ/titlar/Swensk Ordabok>

Johansson, Monica 1997: *Lexicon Lincopense. En studie i lexikografisk tradition och svenskt språk vid 1600-talets mitt*. (Meijerbergs arkiv för svensk ordforskning 21.) Göteborg.

Ohlsson, Stig Örjan 1992: *Språkforskaren Urban Hiärne. Jämförande studier mot europeisk och skandinavisk bakgrund*. Lund: Ambla förlag.

Ralph, Bo 1992: The older dictionaries as sources for Nordic language history. I: *The Nordic Languages and Modern Linguistics* 7. Proceedings of the 7:th International Conference of Nordic and General Linguistics i Tórshavn, 7–11 August 1989. Vol. II. (Annales Societatis Scientiarum Færoensis. Supplementum XVII.) Tórshavn, 493–509.

Ralph, Bo 2000: Svensk lexikografitradition. I: *LexicoNordica* 7, 5–22.

Rogström, Lena u.a: Etablering eller förankring – hur kan man se på arbetet med ordförrådet i 1734 års lag? Föredrag vid Svenska språkets historia 11, 23–24 april 2010 i Uppsala.

Schenberg, Petrus 1739: *Lexicon latino–svecanum*. Norcopiæ.

Swedberg, Jesper 1716: *Schibboleth. Swenska språkets rycht och richtighet*. Skara.

Swedberg, Jesper 1722: *En kortt Swensk Grammatica*. Stockholm.

Anna Helga Hannesdóttir
fil. dr, universitetslektor
Institutionen för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
anna.hannesdottir@svenska.gu.se

En deskriptiv finsk frasordbok

Riina Klemettinen

Muikku-Werner, Pirkko – Jantunen, Jarmo Harri – Kokko, Ossi: *Suurella sydämellä ihan sikana. Suomen kielen kuvaileva fraasisanakirja*[1]. Helsingfors: Gummerus 2008. 467 sidor, ca 4 000 fraser.

1. Introduktion

År 2008 publicerades en finsk frasordbok sammanställd av Pirkko Muikku-Werner, Jarmo Harri Jantunen och Ossi Kokko. Förläggare för *Suurella sydämellä ihan sikana. Suomen kielen kuvaileva fraasisanakirja* (i fortsättningen SSIS) är Gummerus. Det har funnits stort behov för en enspråkig finsk frasordbok, för resultaten av tidigare försök att sammanställa frasordböcker har varit rätt blygsamma. Existerande verk är *Suomalainen fraasisanakirja* (”Finsk frasordbok”) av Sakari Virkkunen (1974), *Naulan kantaan: nykysuomen idiomisanakirja* (”Huvudet på spiken: nufinsk idiomordbok”) av Erkki Kari (1993) och *Aasinsilta ajan hermolla* (”En åsnebrygga i tiden” av Jukka Parkkinen (2005).

SSIS är avsedd att vara en bok både att slå upp i och att läsa i. Enligt baksidestexten ska ordboken vara ett verktyg för alla som skriver eller talar finska. Användargrupper som nämns är författare, redaktörer, översättare, lärare och studenter. Dessutom ska ordboken kunna fungera som hjälpmedel för språkinlärare. Som framgår redan av titeln är SSIS en deskriptiv ordbok, och författarna poängterar i förordet att syftet inte är att normera frasan-

1 Titeln är en kombination av två idiom, där den första delen betyder ’med stort hjärta’ och den andra ’i stora mängder’.

vändningen. I vissa fall ges dock anvisningar om stilvärdet (t.ex. *vardagligt, mycket vardagligt*) och ibland också uppgifter om typiska kontexter (t.ex. *i sporttexter, vanligt i politiska texter*).

SSIS anför nästan 4 000 fraser. Materialet är hämtat ur Språkbanken i Finland och ur tidningar och tidskrifter. Frasurvalet är gjort med modern finska för ögonen och kravet har varit att det ska handla om fullt levande språkmaterial och inte om exempel som bara hittas i ordböcker. När man bläddrar i ordboken konstaterar man att redaktionen har lyckats väl härvidlag. Man hittar sådana modeuttryck som *elämä on* 'sån't är livet' och *aikuisten oikeesti* 'på riktigt, på allvar' (ordagrant 'de vuxnas på riktigt'). Dagsfärskheten har dock sitt pris; det är omöjligt att veta om modeuttryck av det här slaget kommer att etablera sig i språket eller om det handlar om dagsländor.

Frasmaterialet är mångsidigt. Ordboken tar upp både verbfraser (*vaihtaa hiippakuntaa* 'dö', *ottaa hatkat* 'sticka, pysa, ge sig av' och verblösa fraser (*punainen lanka* 'den röda tråden', *juurta jaksain* 'in i minsta detalj'). Den redovisar jämförelser (*päissään kuin ellun kana* 'full som en alika', *kuin hangon keksi* 'som solen i Karlstad', dvs. en liknelse för någon som ler brett), utrop (*istu ja pala* 'har man hört på maken') och olika andra helheter (*siinä paha missä mainitaan* 'när man talar om trollen'). Bevingade ord och ordspråk har författarna enligt egen uppgift exkluderat, men påståendet håller inte riktigt streck, för det ingår en del ordspråksartat material. Ett sådant är *kahden kauppa on kolmannen korvapuusti* vars betydelseinnehåll är att det som är bra för två kan vara till nackdel för en tredje (ordagrant 'en affär mellan två är en örfil för den tredje'). Vidare hittar man exempelvis *se parhaiten nauraa joka viimeksi nauraa* 'skrattar bäst som skrattar sist' och *ei ole koiraa karvoihin katsominen* 'man ska inte döma hunden efter håren'.

2. Frasbegreppet i ordboken

Författarna har valt den samlande benämningen *fraasi* 'fras' för de typer av uttryck som redovisas i ordboken, och som synonym använder de *sanonta* 'fras, talesätt'. De uppger sig vilja undvika termen *idiomi* 'idiom' eftersom de menar att den är för snäv, men trots det förekommer *idiom* på ett flertal ställen i användaranvisningarna. Vad *fras* och *idiom* egentligen betyder kan förstås diskuteras i det oändliga och man kan ha olika mening om vilket slags språkmaterial en ordbok som SSIS borde innehålla. På det hela taget är urvalet dock lyckat.

En aspekt av frasurvalet måste dock lyftas fram. SSIS innehåller en hel del "uppslagsfraser" som består av ett enda ord. Enligt författarna rör det sig om "modeord som ofta används på ett frasartat sätt". En del av sammansättningarna kunde accepteras till nöds, t.ex. *teerenpeli* 'flirtande' (ordagrant 'orrspel'), *maanantaikappale* 'måndagsexemplar' och *ruskeakielinen* 'iställsam'. Orden är idiomartade i den bemärkelsen att deras betydelse inte kan räknas ut utifrån delarnas betydelse. Dock går sammansättningar av det här slaget mestadels att hitta i gängse ordböcker.

Också ett och annat osammansatt ord har tagits med. Det rör sig främst om nyord: *friikki* 'freak', *grillata* 'grilla' och *viileä* 'cool'. Den bildliga betydelsen av *grillata* ('grilla, pressa, ansätta') finns t.ex. inte med i den finska definitionsordboken *Kielitoimiston sanakirja* (2008). Det är dock svårt att förstå varför det i en frasordbok finns med ettordsuttryck som *juurtua* 'få fotfäste' (till *juuri* 'rot'), *venyä* 'klara, prestera; vara flexibel' (konkr. 'töjas ut'), *tulva* 'ström, störtflod' (konkr. 'översvämning') eller *kuilu* 'klyfta'. Den bildliga betydelsen för dessa anges redan i *Suomen kielen perussanakirja* (1990–1994) och till och med i *Nykysuomen sanakirja* (1951–1961), så några nyheter i språket är det inte fråga om. Enda orsaken till att den här typen av ord tagits med verkar vara att

de har en bildlig betydelse; författarna har alltså satt likhetstecken mellan bildlig användning och fras. En sådan utgångspunkt är problematisk med tanke på materialavgränsningen. Språk innehåller mängder av ord som kan användas bildligt och det är omöjligt att göra ett adekvat urval för en frasordbok – och till yttermera visso skulle en ordbok innehållande en mängd ”ettordsfraser” inte längre göra sig förtjänt av benämningen *frasordbok*.

Användaren kan för sin del inte veta varför den bildliga användningen av vissa ord tas upp i ordboken medan man får leta förgäves efter andra.

3. Ordbokens upplägg

Upplägget i ordboken följer traditionen för idiomordböcker: det finns en alfabetiskt ordnad lemmalista och fraserna kommer upp under den första fraskomponenten med bildlig betydelse. Frasen *toinen ääni kellossa* ’annat ljud i skällan’ anförs exempelvis under lemmat *ääni* ’ljud’. Den här principen redovisas i användaranvisningarna, vilket naturligtvis är bra. I praktiken är det dock enklast att söka fraser med hjälp av idiomordbokens förträffliga register.

Ett omfattande och mycket användbart register hör nämligen till ordbokens verkliga förtjänster. Varje innehållsord i de fraser ordboken behandlar återfinns som stickord i registret. Under varje stickord listas samtliga de fraser som ordet ingår i med det lemma som frasen ska sökas under fetstilsmarkerat.

Artiklarna under varje enskilt lemma är uppställda enligt följande. Uppslagsfrasen i fetstil följs av en betydelseförklaring och denna följs i sin tur av mer eller mindre autentiska språkexempel markerade med ett piltecken (▶). Om betydelserna är fler är artikeln indelad i numrerade betydelsemoment. I förklaringsdelen nämns också ord som tenderar att förekomma i kombination med frasen, dvs. kollokaten. För varje betydelse ges i regel två eller tre

exempel, ibland fler. Sist i varje moment kan det finnas uppgifter om stilvärde. I de fall när frasen innehåller egennamn ges också bakgrunden till uttrycket.

4. Frasernas grundform

En stor utmaning för författarna till en idiomordbok är frågan om i vilken form fraserna ska anföras. Idiom brukar presenteras på metanivå som ett slags schema. Denna schemaartade form går under olika benämningar i litteraturen: *kanonisk form, grundform, standardform* osv. (Sköldberg 2004). Jag kommer i det följande att använda *grundform*.

I en idiomordbok vore det på det hela taget rekommendabelt att anföra idiomen med så kompletta valensuppgifter som möjligt. Det är dock inte SSIS linje. Tvärtom leder det avskalade presentationssättet i många fall till tolkningssvårigheter. Som exempel får frasen *olla heikkona* 'vara svag' tjäna. Den betydelse som anges är 'vara förtjust, inte kunna motstå något'. Utan den illativbestämmelse som anger föremålet för känslan kommer frasen dock att avse ett kroppsligt svaghetstillstånd, dvs. att en person är i dåligt skick. Grundformen borde alltså vara *olla heikkona* ***johonkin/johonkuhun*** 'vara svag för något/någon'.

I en del fall är slutresultatet närmast komiskt som följd av att valenselement utelämnats ur de angivna grundformerna. När grundformen anges vara *yllättää housut kintuissa* ('överraska med byxorna nere') kan man som läsare börja undra om det är den som överraskar eller den överraskade som står med byxorna nere. Likaledes undrar man inför frasen *kasata tulisia hiiliä pään päälle* ('samla glödande kol på huvudet') vem som får kolet på huvudet. De angivna formerna borde lyda *yllättää* ***joku*** *housut kintuissa* och *kasata tulisia hiiliä* ***jonkun*** *pään päälle.*

De grundformer som anges i SSIS är oftast så avskalade att de

inte som sådana längre utgör hela fraser utan snarast något slags lexikala eller semantiska kärnor (se Heinonen 2009). Samma kärna kan förekomma i flera fraskonstruktioner, men läsaren måste själv kunna abstrahera de olika fraserna utifrån de exempel som anförs i artiklarna. För den angivna grundformen *olla ilmaa* ('vara luft') ges exempelvis två betydelser: 'det finns obefogat mycket av något' och 'vara fullständigt betydelselös'. I själva verket ansluter sig de här betydelserna till två olika konstruktioner: *jossakin on ilmaa* 'det finns luft i något' (t.ex. i priser, om priserna är obefogat höga) och *joku/jokin on jollekulle ilmaa* 'någon/något är som luft för någon'. Vanligare är i alla fall att den rudimentära kärnan hör samman med en enda konstruktion men att den presenterade grundformen är ofullständig. Exempel på det är *kana kynittävänä/kynimättä* 'en höna att plocka/oplockad'. För att en meningsfull fras ska bildas borde kärnan vara omgiven av lämpliga valenselement: ***jollakulla*** *on kana kynittävänä/kynimättä* ***jonkun kanssa*** 'någon har en höna [sv. gås] oplockad med någon'.

Alla de konstituenter som anförs i grundformen och som hör till konstruktionen på begreppsnivå behöver inte nödvändigtvis alltid realiseras i faktiskt språkbruk, men grundformens funktion är att beskriva den logisk-semantiska strukturen i en situation. Exemplen kan sedan belysa hur det ser ut i faktiskt språkbruk.

I många fall ser grundformen hos fraserna ut som infinitivkonstruktioner: *viitata kintaalla* 'vinka med kalla handen', *vetää välistä* 'sko sig', *panna parastaan* 'göra sitt bästa' osv. Ofta fungerar detta problemfritt, dvs. när subjektet i finit användning tar agentroll. Men i en del fall anförs fraser i infinitivform i SSIS på ett sätt som inte stämmer överens med modersmålstalares språkkänsla: *olla luurankoja kaapissa* (ordagrant 'vara skelett i garderoben [eg. skåpet]'), *olla otsaa* ('vara panna'), *nousta häly* ('bli oväsen'), *tulla pannukakku* ('bli pannkaka'). Problemet med de två första fraserna är att de är habitiva och att ägarrollen därför borde komma till synes: ***jollakulla*** *on luurankoja kaapissa* 'någon har skelett i

garderoben', ***jollakulla*** *on otsaa (tehdä jotakin)* 'någon har panna (att göra något)'. I de två senare saknas temarollen: ***jostakin*** *nousee häly* 'det blir uppståndelse kring något' och ***jostakin*** *tulee pannukakku* 'det blir pannkaka av något'.

En annan viktig aspekt vid infinitivkonstruktioner är oppositionen mellan animat och icke-animat subjektreferent. I finskans så kallade *a*-infinitiv har subjektet en stark tendens att vara personsyftande, och som finskspråkig tänker man sig i första hand ett personsubjekt för *a*-infinitiver där ett sådant är möjligt (Visapää 2008). Fraser som *kasvattaa luonnetta* och *tehdä pahaa/huonoa* är därför problematiska som grundformer. Bägge fraserna avser fall med inanimat subjekt: ***jokin*** *kasvattaa luonnetta* 'något danar karaktären' och ***jokin*** *tekee pahaa* 'något känns illa'. Kombinationen *tehdä pahaa* 'göra illa' kan användas också med mänskligt subjekt men då handlar det inte om ett idiom och betydelsen är en annan. I all synnerhet i de fall då flera fraser med olika slag av subjektreferent hamnar efter varandra i artiklarna vore utförligare angivelser av nöden. Ett exempel på det är artikeln *mennä* med följande fraser i en följd: *mennä asiaan* '[gå] till saken', *mennä helppoon* 'låta sig luras', *mennä nappiin* 'lyckas, klaffa' (till *nappi* 'knapp') och *mennä putkeen* 'lyckas' (till *putki* 'rör'). De två första fraserna tar personsubjekt, de två senare saksubjekt, men för alla fyra läser man först spontant in ett personsubjekt till *mennä* 'gå, fara'.

5. Variation

Traditionellt har idiom ansetts vara mer eller mindre invariabla men numera är uppfattningen den att det i faktiskt språkbruk förekommer en hel del variation (se t.ex. Clausén 1993 och Sköldberg 2004). Vissa idiom medger större modifikationer än andra. Vid sidan av aktualiteten bär SSIS prägel av en uttrycklig strävan att framhålla variationsmöjligheterna. Det här är inte någon helt

oproblematisk utgångspunkt, eftersom en ordbok i regel förväntas åskådliggöra typiska fall eller det generella i språket. Det vore således viktigt att ta ställning till vilka typer av variation det över huvud taget lönar sig att redovisa i en ordbok och hur denna variation ska redovisas (Clausén 1993).

Grundformen för en fras är alltså ett slags schema eller en generalisering, som i ordböcker och andra metatexter utformas utifrån den variation som uppträder i språkbruket. Det säger sig självt att alla variationsmöjligheter inte går att täcka in på generell nivå. Om den angivna grundformen inte tillåts innehålla något annat än de element som är gemensamma för alla upptänkliga varianter av en viss fras blir resultatet lätt att frasangivelsen blir alltför knapphändig och ibland rentav obegriplig. Detta är fallet i SSIS.

I vissa fall förefaller det som om grundformerna presenteras i ofullständig form på grund av att verbet kan variera: om flera verb är tänkbara i en viss struktur redovisas ibland inget verb alls. Fraser som *eri sarjassa* och *matto alta* ('i olika serier' resp. 'mattan undan') skulle obetingat kräva verb för att bli meningsfulla. På finska säger man vanligtvis *painia eri sarjassa* (sv. 'stå i en klass för sig' – ordagrant 'brottas i en annan serie') och *vetää matto jonkun alta* 'rycka undan mattan för någon'. Det räcker inte att de typiska verben anges som kollokat, för de stympade grundformerna fungerar inte som sådana. Ett bättre alternativ vore att anföra frasen med det mest typiska verbet och ange att variation kan förekomma. En annan möjlighet vore att ange flera olika verb i grundformen åtskilda med snedstreck. Det senare alternativet har ställvis utnyttjats i SSIS, t.ex. frasen *juoda/niellä/maistaa katkeraa/karvasta kalkkia* ('dricka/svälja/smaka den bittra/beska kalken').

Också andra element än verb kan vara utelämnade ur grundformerna trots att de är nödvändiga för helheten. Ett exempel på det är frasen *kehittyä askelin* 'utvecklas i steg'. Frasen kräver en angivelse av stegens art för att bli begriplig. Ett attribut till *askelin* borde alltså ingå. De mest typiska bestämningarna är *suurin* och

pienin ('stora' resp. 'små') och dessa borde ha redovisats. Av språkbruksexemplen kunde det ha framgått att adjektivet är utbytbart mot vissa andra t.ex. *mullistavin* 'revolutionerande' eller *pikkiriikkisin* 'pyttesmå'.

6. Betydelseangivelserna

På grund av att fraserna presenteras i så avskalad form i SSIS kommer betydelseangivelserna ofta att halta. För nominalfrasen *vaaleanpunaiset silmälasit* 'rosa glasögon' ges förklaringen 'se saker och ting som bättre än de är, som positiva'. Att ha en verbfras som förklaring till en nominalfras är inkongruent. Det ovan nämnda *matto alta* 'mattan undan' utan angivet verb anges betyda 'beröva [ngn] möjligheter'. Bristen på symmetri beror på att betydelseangivelsen gäller den kompletta fraskonstruktionen medan frasen finns presenterad i stympad form. Obalansen skulle undvikas om fraserna angavs i sin helhet: *nähdä jokin asia vaaleanpunaisten silmälasien läpi* 'se på livet (eg. något) genom rosa glasögon' och *vetää matto jonkun (jalkojen) alta* 'dra undan mattan för någon/ dra mattan undan fötterna på någon'). Ett annat alternativ vore att förklara fraserna med hjälp av definitioner av COBUILD-typ: "en person som ser på livet genom rosa glasögon har en orealistiskt positiv uppfattning och bortser från de svårigheter som kan vara förknippade med en sak".

Ytterligare exempel på fraser med osymmetriska förklaringar är *langat käsissä* ('trådarna i händerna') som förklaras med verbet *hallita* 'behärska' (jfr *hålla i trådarna*), nominalfrasen *kypsä ikä* 'mogen ålder' förklarad med adjektiv/particip *vanha, vanheneva* 'gammal, åldrande', *Pandoran lipas* 'Pandoras ask' med betydelseangivelsen 'oväntad aktion, som ofta har negativa följder', *terävä pää* 'skarp hjärna' (eg. 'skarpt huvud') med förklaringen *paljon älyä* 'mycket intelligens'.

Det förekommer också ofta osymmetri i form av perspektivbyte mellan fras och förklaring. Frasen *pyöriä kielen päällä* (ordagrant 'snurra på tungan') förklaras t.ex. med 'vara på väg att komma på, nästan minnas'. I förklaringen syftar det utelämnade frassubjektet på personen som försöker komma på en sak medan frassubjektet i själva frasen skulle avse den sak som personen inte kan dra sig till minnes. Detsamma gäller *nousta/mennä päähän* 'stiga åt huvudet'. Frasen har två förklaringar 'bli för stolt över sina prestationer' och 'bli berusad'. Subjektreferenten för själva frasen avser den omständighet som föranleder stoltheten respektive den dryck som åstadkommer berusningen, medan det tänkta subjektet i förklaringarna åsyftar den stolta respektive berusade personen. Grundformen borde alltså lyda ***jokin*** *nousee/menee* ***jollakulla/jollekulle*** *päähän* 'något stiger/går någon åt huvudet'. Förklaringarna kunde då ges i elativform *ylpistymisestä* ('om att bli stolt') och *humaltumisesta* ('om att bli berusad'), som är en hävdvunnen förklaringstyp i finska ordböcker.

För exempelvis frasen *saada tietoonsa* 'få reda på' har de två förklaringsdelarna sinsemellan olika perspektiv: 'få veta' och 'bli känd' ('**någon** får veta något' men '**något** blir känt').

Många fraser är försedda med så knapphändiga betydelseangivelser att de inte fungerar som förklaringar: *sivu suun* förklaras med 'förbi'. (Frasen borde lyda *jokin menee joltakulta sivu suun* 'något går någons näsa förbi' [ordagrant 'förbi munnen']). En så lakonisk förklaring säger inget om den kontext frasen används i, och ofta är konnotationerna lika viktiga som den denotativa betydelsen när en fras och dess användning ska förklaras. Också i det här fallet vore en förklaring av COBUILD-typ beaktansvärd: "om man inte lyckas få något som man varit nära att få säger man att det går ens näsa förbi". Uttrycket *naisen aseet* 'en kvinnas vapen' förklaras föga uttömmande med 'utvägar som kvinnor använder sig av'. Ordboksanvändaren måste utifrån sin kunskap om världen försöka räkna ut vad det kan tänkas handla om för utvägar. Lika

kryptiskt blir det när *parempiin suihin* förklaras med 'äten'. Tanken bakom uttrycket *kadota/mennä parempiin suihin* (ordagrant 'försvinna i bättre munnar') är att någon äter eller dricker upp något som någon annan bespetsat sig på.

7. Exemplen

Exemplen i SSIS är till största delen autentiska. Vissa smärre modifieringar kan ha gjorts för begriplighetens skull och i någon mån har exemplen också förkortats. Att hitta lämpliga exempel i ett korpusmaterial kan vara överraskande besvärligt (Lorentzen 1999), och SSIS är inget undantag därvidlag. Vissa av exemplen är rentav mer förvirrande än klargörande i relation till de fraser de ska illustrera. För den ovan nämnda frasen *naisen aseet* anförs två exempel. Det ena är *menestyvä urheilujoukkue tarvitsee naisen aseet* 'ett framgångsrikt idrottslag behöver kvinnliga vapen' och det andra lyder *rinnat toimivat naisen aseena, toisaalta hänen turvanaan* 'brösten är en kvinnas vapen och också hennes trygghet'. Inget av exemplen har någon som helst funktion när det gäller att åskådliggöra uttryckets betydelse.

Det hade också varit på sin plats att gallra ut egennamn ur exemplen. Personer som Aigars Cipruss (ishockeyspelare) eller Green Day (ett punkband) är knappast särskilt välkända ens för dagens ordboksanvändare och än mindre för användare om fem till tio år. I många fall kan egennamn rikta uppmärksamheten på ovidkommande saker i stället för på själva idiomet. Under *kuin yö ja päivä* ('som natt och dag') återfinns det autentiska exemplet *Pyhiksen ja Karhun pallojen ero on kuin yöllä ja päivällä* 'skillnaden mellan Pyhis och Karhus bollar är som natt och dag'.

Också andra element som snarare väcker frågor än ger svar kunde med fördel ha utgått ur exemplen. När man som ordboksanvändare stöter på exemplet *toisinaan siinä onnistutaan nautit-*

tavasti, joskus menee pahasti överiksi 'ibland lyckas det bra, ibland slår det över' undrar man lätt vad det är som har varit lyckosamt (frasen *mennä överiksi* 'slå över, gå till överdrift'), och inför *samat sanat, mekin naurettiin sille* 'samma här, vi skrattade också åt det' undrar man bara vad skrattet har gällt (frasen *samat sanat)*.

8. Till slut

Idiomordboken SSIS förutsätter en hel del god vilja från användarnas sida. För att använda ordboken ska man egentligen helst redan känna till de redovisade fraserna och veta hur de används och vad de betyder. De anförda grundformerna är ofta alltför knapphändiga medan betydelsebeskrivningarna är vida och diffusa. Ordboken är inget verktyg för den som vill kontrollera den adekvata användningen av en viss fras.

Att den skulle vara ett bra hjälpmedel för språkinlärare är definitivt för mycket lovat med tanke på de avskalade fraserna och de rudimentära och oklara betydelseangivelserna.

Till bokens förtjänster hör ett mångsidigt frasmaterial, god aktualitet och ett utmärkt register.

Litteratur

Ordböcker

Kari, Erkki 1993: *Naulan kantaan: nykysuomen idiomisanakirja.* Helsinki: Otava.

Kielitoimiston sanakirja (cd-rom) ('Språkbyråns ordbok'). Helsinki: Kotimaisten kielten tutkimuskeskus ja Kielikone Oy 2008.

Nykysuomen sanakirja ('Nufinsk ordbok'). Valtion toimeksiannosta teettänyt Suomalaisen kirjallisuuden seura. Helsinki: WSOY 1951–1961.

Parkkinen, Jukka 2005: *Aasinsilta ajan hermolla.* Helsinki: WSOY.
Suomen kielen perussanakirja (’Finsk basordbok’). Helsinki: Kotimaisten kielten tutkimuskeskus ja Painatuskeskus 1990–1994.
Virkkunen, Sakari 1975: *Suomalainen fraasisanakirja.* Helsinki: Kustannusosakeyhtiö Otava.

Annan litteratur

Clausén, Ulla 1993: Idiom och variation. I: *Nordiska studier i leksikografi 2*. Rapport fra Konference om Leksikografi i Norden 11.–14. maj 1993, 47–52.
Heinonen, Tarja 2009: Idiom, liknelser och kollokationer i två finska ordböcker. I: *LexicoNordica 16*, 141–159.
Lorentzen, Henrik 1999: Jagten på det gode citat. Om vanskelighederne ved at finde egnede ordbogseksempler i et korpus. I: *Nordiska studier i lexikografi 5*. Rapport från Konferens om lexikografi i Norden Göteborg 26–29 maj 1999. Göteborg: Göteborgs universitet, 202–216.
Sköldberg, Emma 2004: *Korten på bordet: innehålls- och uttrycksmässig variation hos svenska idiom.* Göteborg: Meijerbergs arkiv för svensk ordforskning.
Visapää, Laura 2008: *Infinitiivi ja sen infiniittisyys. Tutkimus suomen kielen itsenäisistä A-infinitiivikonstruktioista*. Helsinki: SKS.

Riina Klemettinen
forskare, ordboksredaktör
Forskningscentralen för de inhemska språken
FI-00500 Helsingfors Finland
riina.klemettinen@kotus.fi

Översättning från finska: Nina Martola

Svensk ordbok – en guldgruva för språkintresserade

Kristina Nikula

Svensk ordbok utgiven av Svenska Akademien, Band 1 (A–L), band 2 (M–Ö) i en box. Utarbetad vid Redaktionen för Svenska Akademiens samtidsordböcker, Lexikaliska institutet, Institutionen för svenska språket vid Göteborgs universitet. Förlag: Norstedts Akademiska förlag (i distribution). Stockholm 2009.

0. Inledning

Hösten 2009 utkom *Svensk ordbok* (SO) som bär samma namn som *Svensk ordbok* (SOB) från 1986 men som i själva verket ska ses som en uppdaterad och utvidgad version av *Nationalencyklopedins ordbok* (NEO 1995–1996). NEO var i sin tur en reviderad och utvidgad utgåva av SOB. SO, som består av två band, band 1 (A–L) och band 2 (M–Ö), uppges innehålla ca 65 000 uppslagsord[1] som alla är försedda med betydelsebeskrivning. Med betydelsebeskrivningar till betydelsenyanser och idiom medräknade uppskattas betydelsebeskrivningarna till drygt 100 000. Varje artikel innehåller uppgifter om ordklasstillhörighet, böjning och uttal. För ord som inte antas bereda uttalssvårigheter och för sammansättningar anges bara accent. De formella uppgifterna avslutas med information om ordled. Den första definitionen och övriga numrerade definitioner anger huvudbetydelse, som i många fall

1 Antalet lemman anges i företalet till ca 65 000 (SO:VII) och i inledningen till 64 500 (SO:IX). En uppskattning av antalet lemman med utgångspunkt i en ”normalsida” skulle säkerligen inte ge en mera rättvisande uppskattning.

kompletteras med betydelsenyanser på egen rad som föregås av en liten cirkel. Betydelsebeskrivningen följs vanligen av exempel på användningen. Vissa artiklar innehåller därtill uppgifter som stilvärde, fackområde o.d. I SO ingår också kollokationer och idiom samt uppgifter om valens. Varje artikel avslutas med årtal för första belägg samt ordets etymologi och allra sist förekommer på sina ställen citat och en ruta som behandlar språkriktighetsfrågor i anslutning till uppslagsordet. Det är med andra ord en ansenlig mängd information som ges på de 3 736 sidorna. En överskådlig presentation av innehållet i ordboksartiklarna finns på den främre pärmens insida i vartdera ordboksbandet.

1. Ordbokens syfte och användare

SO är utrustad med ett förord, ett företal och en inledning. I förordet avgränsas SO mot SAOB, i företalet presenteras bl.a. SO:s förhistoria kort och omnämns den empiriska bas, *Lexikalisk databas*, som ordboken bygger på.

Alla tre inledande texter uppger på sitt sätt syftet med ordboken. Sammantaget strävar SO efter att ge så fullständiga upplysningar som möjligt om alla viktiga ord i ordförrådet i modern svenska. Tyngdpunkten uppges vara lagd på ordens betydelse och användning. Ordboken vänder sig till alla som vill lära sig mer om det svenska ordförrådet och ska vara en hjälpreda vid reception och produktion av text samt ge en uppfattning om ordens härkomst (SO:V, VII, IX).

Eftersom SO uppges vända sig till dem som vill lära sig mer om det svenska ordförrådet (SO:IX), kan användaren vara i princip vem som helst, både användare med svenska som modersmål och i våra dagar allt fler användare med ett annat modersmål än svenska. SO borde därtill också kunna användas i vetenskapligt syfte. SO:s funktion som både receptions- och produktionsordbok stäl-

ler därtill stora och delvis motstridiga krav på ordboken vad gäller struktur och tillgänglighet. Det är mot bakgrund av beskrivningen av SO:s funktion som jag i huvudsak kommer att bedöma hur väl man lyckats nå målet. I fortsättningen tar jag upp och diskuterar fenomen som jag funnit intressanta och/eller problematiska. Framställningen följer i stort ordboksartiklarnas uppläggning.

2. Databas och lemmaselektion

SO baserar sig på *Lexikalisk databas* som utarbetats vid Lexikaliska institutet vid Göteborgs universitet. Hur omfattande denna databas är och vilka typer av text den baserar sig på framgår emellertid inte. Huruvida källmaterialet är baserat på talat och/eller skrivet språk och balanserat med hänsyn till olika typer av text framgår sålunda inte heller. Detta gör att man endast kan anta att källmaterialet till databasen är någorlunda representativt för den moderna svenskan i alla dess varianter; den perfekta korpusen finns helt enkelt inte, därtill är språk alltför dynamiska (Atkins & Rundell 2008:95). För ordboksanvändaren i gemen är bristen på information säkerligen av underordnad betydelse men ur lingvistens synvinkel otillfredsställande då det inte heller ges någon hänvisning till källor som kunde ge information. Frågan aktualiseras inte minst som det råder en viss oenighet om vilka typer av text som ger det lemmaurval som bäst motsvarar användarens behov (Engelberg & Lemnitzer 2009:89).

Enligt företalet innehåller SO ”alla viktiga ord i det svenska ordförrådet” (SO:IX). Eftersom påståendet är föga anspråkslöst inställer sig omedelbart frågan på vilka grunder de viktiga orden har valts. Man får anta att detta som brukligt skett på basis av frekvens, även om det vore viktigare att de ovanliga orden fanns med därför att det är de som i första hand kan förväntas förorsaka ordboksanvändaren problem (Svensén 2004:80–82).

En ordbok kan aldrig vara heltäckande. En ordbok behöver inte heller vara heltäckande om man med detta avser att ”varje” svenskt ord ska förekomma som uppslagsord. Uppgifter om enligt vilka principer lemmaselektionen har genomförts är därför viktiga med tanke på användaren. I SO uppges genomskinliga sammansättningar ha tagits med enbart i begränsad utsträckning (SO:IX). Det är därför naturligt att t.ex. *bordsduk* saknas som lemma. Den som tvivlar på att ordet existerar eller vill veta om sammansättningsfogen innehåller ett *-s-*, hittar det däremot bland de morfologiska exemplen under ordet *bord* i formen *bord(s)duk* och under *duk* som *bordsduk*. På motsvarande sätt hittar man *minkpäls* under såväl *mink* som *päls* och *skjortkrage* under både *skjorta* och *krage*. Man måste likväl utgå ifrån att antalet genomskinliga sammansättningar som ges i ordboken är begränsat och att de ingalunda alltid finns anförda under både för- och efterled. Men även om placeringen verkar slumpmässig och därigenom inte speciellt användarvänlig är det, bl.a. med tanke på sammansättningsfogen, ändå att föredra framför att de genomskinliga sammansättningarna helt skulle ha uteslutits ur ordboken.

Enbart en bråkdel av alla tänkbara facktermer uppges ha tagits med i ordboken. För att en fackterm ska uppföras i en allmänspråklig ordbok bör denna ingå i texter som vänder sig till icke-specialister. Fackområden som det kan finnas skäl att observera är enligt Svensén (2004:89f.) bl.a. konsumtion av varor och tjänster (t.ex. medicin) och områden som under en viss tid väcker stor uppmärksamhet i media. En granskning av beteckningar för sjukdomar visar att bl.a. *alzheimer*, *asiaten*, *fågelinfluensa*, *galna ko-sjukan*, *ms* e. *Ms*, *sars*, *Parkinson* e. *Parkinsons sjukdom*, *psoriasis*, *scrapie* och *vinterkräksjuka* finns uppförda förutom en massa andra ”vanliga” åkommor som *influensa*, *kikhosta*, *röda hund, vattkoppor* och *mässling* med flera. Den så kallade *Kumlingesjukan* eller *TBE* (ett slags hjärnhinneinflammation) finns däremot inte uppförd, kanske därför att sjukdomen hittills inte utgjort samma

problem och blivit lika omskriven i Sverige som i Finland. Fältet SJUKDOM förefaller trots detta ha god täckning.

En kontroll i SO av ordförrådet i en artikel i Svenska Dagbladet (9.12.2009) om budgetkrisens Grekland visade att av de ekonomiska termer som kunde tänkas skapa problem oväntat många finns anförda (t.ex. *kreditbetyg*, *riskpremie*). Däremot inte *EU*, som också nämns i texten, även om initialord finns anförda. Förklaringen till detta är att proprier normalt inte brukar anföras i enspråkiga ordböcker (Svensén 2004:91–93).

Det verkar som om man i SO lyckats få med det mest centrala i det svenska ordförrådet, såväl högtidligt som vardagligt. Här finns allt från *nejd* <högt.> till *snackis* <vard.> Här finns *bonusbarn*, *mess*, *mp3-spelare*, *platt-tv* och *självmordsbombare*. Också initialord och förkortningar finns anförda. Ett enstaka finlandssvenskt ord har jag påträffat: *råddig* 'oklar och virrig', men detta markeras inte som finlandssvenskt utan som provinsialism. Av 327 finlandssvenska ord i SAOL Plus har Martola (2010:199f.) påträffat endast åtta i SO, däribland *råddig*. Över lag förefaller SO ställa sig ganska kallsinnig till provinsialismer trots att de är viktiga ur receptionssynvinkel och inte heller är helt oviktiga med tanke på produktion av text.

3. Betydelse

Betydelsebeskrivningen är central i en enspråkig betydelseordbok som SO. Undersökningar[2] har visat att användarna oberoende av modersmål oftast anlitar enspråkiga ordböcker för att få reda på ords betydelse. (Engelberg & Lemnitzer 2009:87.) Med hänsyn till användarens behov kan betydelsebeskrivningen ges olika innehåll

2 En stor del av undersökningarna från 1980- och 1990-talet har dock underkänts till följd av att de inte ansetts fylla de metodiska minimikraven (Engelberg & Lemnitzer 2009:85f.).

och form: bland annat krävs att innehållet i ordboksartiklarna presenteras på ett språk som motsvarar användarens lingvistiska kompetens, inte kräver specialkunskaper eller tvingar användarna att lära sig ovanliga lexikografiska konventioner. (Atkins & Rundell 2008:407–413, se också Lorentzen & Nimb 2010:333–335.)

3.1. Betydelsebeskrivningen

I allmänspråkliga ordböcker är betydelsebeskrivningarna vanligen intensionella, vilket innebär att genus proximum anges och därtill så många särdrag som är nödvändiga men samtidigt tillräckliga för att avgränsa begreppet mot andra element i klassen. SO utgör i detta avseende inte något undantag (SO XIV). Idealet har länge ansetts vara att i den mån det är möjligt använda ett genus proximum som kräver ett enda särdrag för att nå den önskade avgränsningen av begreppet, t.ex. **rektangel** 'rätvinklig fyrhörning' (Svensén 2004:274). Detta är emellertid svårt att uppnå inom områden med en välutvecklad taxonomi, t.ex. biologi, och långt ifrån varje ord låter sig beskrivas enligt denna enkla modell. I och med utvecklingen av prototypsemantiken och korpuslingvistiken har man därför i stället för att isolera nödvändiga och tillräckliga särdrag börjat sträva efter att visa vad som är normalt snarare än vad som är nödvändigt och att förse användarna med information som hjälper till att identifiera prototypiska medlemmar i en kategori (Atkins & Rundell 2008:276–280, 416–419).

I SO definieras på traditionellt vis *marulk* med genus proximum, *rovfisk*, som i sin tur på en högre nivå i hierarkin definieras av ett annat överbegrepp, *fisk*, och ett antal särdrag som ansetts nödvändiga för att ordboksanvändaren ska kunna identifiera den ifrågavarande fisken:

> **marulk** en bottenlevande rovfisk med stort huvud och en förlängd, utbuktad ryggfenstråle på nosen en sorts vajande

metspö som fisken lockar till sig sitt byte med; förekommande i Skagerrak och Kattegatt {JFR **kotlettfisk**}

Även en användare som inte känner till marulken har tack vare den detaljerade beskrivningen, och till följd av att man appellerar till användarens kunskap om världen, möjlighet att identifiera fisken.

Gränsdragningen mellan språklig och encyklopedisk information och användningen av encyklopedisk information i ordböcker har tidvis diskuterats intensivt. (Se t.ex. Lundbladh 1999, Svensén 2004:353–357, Engelberg & Lemnitzer 2009:7–9) I SO definieras **euro** som 'den gemensamma myntenheten inom den europeiska valutaunionen (EMU)'. En minimal betydelsebeskrivning som **euro** 'en myntenhet' och **marulk** 'en (rov)fisk' kunde förslå om behovet av information uppstått vid reception av text men vid produktion av text är detta inte tillräckligt. En kort beskrivning som 'en myntenhet' resp. '(rov)fisk' kan visserligen tjäna som vägvisare till litteratur som kan ge närmare upplysningar men förlänger samtidigt sökprocessen för den som inte nöjer sig med den knappa beskrivningen utan behöver mera precis information.

Lax beskrivs i SO som 'en större vandrande fisk med långsträckt, spolformig kropp och läckert rosa kött främst förekommande i Nordatlanten'. Att fisken är välsmakande eller en omtyckt matfisk anges också för *gös* och *sik* men inte för t.ex. *gädda* eller *strömming*. Jag håller gärna med om att gäddan inte utmärker sig genom att vara välsmakande men strömmingen vore väl så bra som laxen värd beröm. Tycke och smak ska man inte diskutera och borde därför inte heller, kunde man tycka, ges plats i betydelsebeskrivningar i ordböcker. Det finns ju folk som tycker att **surströmming** 'strömming som är konserverad genom jäsning en norrländsk specialitet', är en delikatess, men varken detta eller lukten, som kanske bäst karakteriserar surströmmingen, nämns i betydelsebeskrivningen. Om man på basis av observerbara data

likväl kunnat konstatera att t.ex. vissa fiskarter oftare än andra åtföljs av positiva konnotationer i fråga om smak, är det naturligt att anföra dessa i betydelsebeskrivningen om de kan bidra till att hjälpa ordboksanvändaren att identifiera fisken (jfr Atkins & Rundell 2008:426f.). **Ros** som beskrivs som 'typ av stor, ofta väldoftande blomma på taggig stjälk' förknippas med en känsla av välbehag till följd av doft och utseende. I de fall konnotationerna är av betydelse för att man ska förstå metaforik är det nödvändigt att de tas med i betydelsebeskrivningen. En metafor som **inte vara ngn dans på rosor** 'inte vara bekymmersfritt' kan vara svår att förstå (Svensén 2004:279f.) eller kan till och med missförstås utan uppgift om konnotationerna, för rosen har också törnen som kunde ge upphov till motsatta konnotationer och metaforer.

Beskrivningen av *svin* väckte kritik då SOB kom ut. Betydelsebeskrivningen 'typ av partåigt hovdjur med lång, kraftig kropp, förlängt nosparti och korta ben' återkom i NEO med tillägget 'ej idisslande'. Samma definition som i SOB återkommer också i SO. Med tanke på folk i allmänhet är denna betydelsebeskrivning inte optimal då man inte kan utgå ifrån att innebörden av vare sig *partåig* eller *hovdjur* är bekant. Det tidigare nämnda *(rov)fisk* ger åtminstone information om att en fisk avses medan *partåig* och *hovdjur* kräver ett visst mått av utomspråklig kunskap. **Dörr** beskrivs i SO som 'vertikal, svängbar skiva för tillslutning av ingångs- eller förbindelseöppning i en byggnad'. Vissa begrepp, t.ex. *svin* och *dörr*, är det i själva verket mer eller mindre omöjligt att beskriva någorlunda enkelt verbalt. I dessa och i många andra fall skulle en illustration därför underlätta förståelsen åtskilligt men då var det ju också frågan om en annan typ av ordbok. Också med hjälp av språkliga exempel kan förståelsen underlättas (se SO **dörr**), men idealet vore att förklaring och exempel var självständiga och kunde förstås oberoende av varandra (Atkins & Rundell 2008:454). Både antonymer och synonymer bidrar därtill på sitt sätt till betydelsebeskrivningen och synonymerna ger dessutom användaren en

möjlighet att variera ordförrådet. SO är över lag generös med information i betydelsebeskrivningen och med exempel. Med tanke på ordbokens dubbla funktion – den ska fungera både vid reception och produktion – och tillfredsställa en synnerligen heterogen målgrupp är detta säkerligen den mest framkomliga vägen.

Oriktiga betydelsebeskrivningar ska enligt Bergenholtz (2005) inte vara sällsynta i ordböcker. Trots ett flitigt bläddrande har jag dock inte lyckats hitta oriktiga betydelsebeskrivningar i SO.

Även om SO i betydelsebeskrivningen uppges följa den traditionella typen med genus proximum och nödvändiga och tillräckliga differentiae specificae i betydelsebeskrivningen verkar man likväl inte vara helt opåverkad av prototypsemantiken som lyfter fram det som är typiskt eller normalt och den vägen hjälper användaren att identifiera referenten.

3.2. Betydelsebeskrivningens struktur

Vid innehållsligt besläktade ord borde man som regel i beskrivningen av betydelsen vara konsekvent vid valet av genus proximum, och bl.a. välja rätt särdrag; ur formell synvinkel borde ord som tillhör samma semantiska fält beskrivas på likartat sätt. (Svensén 2004:277, Atkins & Rundell 2008:392f.) I SO förverkligas detta inte konsekvent i beskrivningarna av t.ex. de tidigare nämnda sjukdomarna. **Alzheimer** beskrivs som 'en sjukdom som drabbar hjärnan och leder till demens'. Hjärnsjukdomar är **galna ko-sjukan** 'en dödlig hjärnsjukdom bland nötkreatur' och **scrapie** 'en dödlig hjärnsjukdom bland får'. Som nervsjukdomar beskrivs **Parkinsons sjukdom** 'en nervsjukdom med skakningar, stelhet och hämmade muskelrörelser' och **ms** 'en kronisk nervsjukdom [...]', medan **psoriasis** är 'en kronisk hudsjukdom med stark avfjällning av hud'. Resten av de ovan nämnda sjukdomarna beskrivs på följande vis: **fågelinfluensa** 'en smittsam virussjukdom som vanligen drabbar tamfåglar men kan överföras till människor',

vinterkräksjuka 'en infektionssjukdom som orsakar illamående, diarréer och kräkningar', **sars** 'typ av smittsam, mycket allvarlig lunginflammation' och **asiaten** 'typ av influensa' som ju också är en virussjukdom. Förutom genus proximum 'en sjukdom' borde i betydelsebeskrivningen av sjukdomar som differentia specifica ingå åtminstone symptom (jfr **vinterkräksjuka**); fakultativa är däremot svårighetsgrad (jfr **sars**), lokalisering (jfr **Alzheimer**), orsak (jfr **fågelinfluensa**) och insjuknad (jfr **scrapie**) (Atkins & Rundell 2008:393). Bäraren av en viss sjukdom anges då det är frågan om en djursjukdom eller om en sjukdom som drabbar både människor och djur, i övriga fall får man utgå ifrån att människan är sjukdomsbärare. I fråga om *Alzheimer* framgår implicit att det är frågan om en sjukdom som drabbar människor eftersom sjukdomen uppges leda till demens, men i fråga om t.ex. *sars* framgår inte om denna sjukdom kan drabba människor och/eller djur. Det förefaller som om betydelsebeskrivningarna i SO motsvarar i stort hur folk i allmänhet skulle beskriva sjukdomar. Strukturering av betydelsebeskrivningarna visar att det är nödvändigt men samtidigt inte problemfritt att kombinera användningen av hierarkiska särdrag med prototyptänkande.

4. Flerordslemman: substantiv, prepositionsuttryck och partikelverb

I SO ingår som uppslagsord också flerordsuttryck som *au pair*, *galna ko-sjukan*, *grand old man* m.fl. substantiv men också prepositionsfraser som *i dag*, *till fullo* och partikelverb som *gå över*, *hitta på*. Även om de flerordsuttryck som inleds exempelvis med *till* är rätt många är de lätta att hitta eftersom det första ordet i den senare delen i kombinationen avgör placeringen, dvs. *till fullo* återfinns under *till* och efter *till freds*. Avsteg från denna enkla regel förekommer bl.a. vid *för övrigt* som inte finns uppfört som lemma

utan måste sökas under *övrigt* och vid *förresten* som också uppges kunna uppträda i formen *för resten* utan att denna form förekommer som uppslagsord.

Också partikelverb är mestadels lätta att hitta i SO även om man inte alltid är konsekvent i fråga om placeringen. Ofta uppträder partikelverben som lemman både med sär- och sammanskrivning, t.ex. *utkomma – komma ut,* andra gånger påträffas de enbart i artiklarna för simplexverben (Martola 2010:195f., 202). I de fall den ena varianten är vanligare än den andra anges detta (SO:XIV). Vad som ska räknas som ett partikelverb kan vara problematiskt men ofta visar den syntaktiska kontexten och i vissa fall i kombination med semantiska faktorer enligt Sundman (2010:314) att prepositionen ska tolkas som en partikel. Av de nitton exempel[3] på partikelverb som Sundman (2010:314) behandlar, återfinns tolv som lemman i SO. De partikelverb som saknas som uppslagsord är *hoppa till* och *marschera på* samt *ta ifrån*, som finns anfört som fast sammansättning **frånta** (ofta lös förb. i formen **ta ifrån**). Vid *dra* (ibl. med partikel, t.ex. **in**, **ner**, **upp**, **ut**), *hälla* (ofta med partikel, särsk. **i**, **på**, **upp**, **över**) och *skrapa* (ofta med partikel, särsk. **bort**) anges i ordboksartikeln att verbet kan förekomma med partikel medan g*å med på* finns anfört enbart som valensuppgift. Över lag är SO ändå rätt generös med partikelverb. (Jfr dock Martola 2010:197.) Efter t.ex. *gå* v. anförs inte mindre än tjugosju partikelverb som lemman, allt ifrån *gå an* till *gå över*. Att partikelverben uppträder som lemman utökar kännbart lemmabeståndet men det är tillfredsställande ur ordboksanvändarens synvinkel att de har en synlig placering och därmed är lätta att upptäcka. Det som till en början verkar vara inkonsekvenser i fråga om placeringen av partikelverben kunde tänkas bero på att de partikelverb som inte anges som lemma har en mera inskränkt användning (jfr an-

3 Följande partikelverb finns anförda hos Sundman: *bryta av, dra för, dra på, dricka ur, gå med på, gå under, hoppa av, hoppa till*, *hålla med, hälla i, koppla ur, marschera på*, *skrapa av*, *skriva om*, *skynda på, sätta på*, *ta med*, *ta ifrån, vara med*.

märkningen *ibl.* med partikel, *ofta* med partikel), något som likväl t.ex. partikelförbindelser med *åt* talar emot, då t.ex. *klämma åt* inte uppförs som lemma (Martola 2010:197f.).

5. Språkliga exempel: sammansättningar, kollokationer och fria illustrationer

SO innehåller gott om språkliga exempel. De morfologiska exemplen består av sammansättningar där uppslagsordet ingår antingen som för- eller efterled: *–handflata*; *–handklappning*; *–handskriven*; *högerhand*. Pilen framför vissa ord anger att sammansättningen också förekommer som uppslagsord. De morfologiska exemplen kan vara till hjälp för den som inte hittat en viss sammansättning som uppslagsord. Även om SO strävar efter att vara användarvänlig måste man kunna förutsätta att användaren, om ett ord saknas som uppslagsord, också letar bland de sammansättningar som anförs under den aktuella sammansättningens för- och/eller efterled.

De syntaktiska exemplen består av kollokationer och/eller friare illustrationer till uppslagsordets användning. Kollokationerna, t.ex. *begå självmord*, *fatta/ta ett beslut*, som är viktiga om man vill använda uppslagsordet optimalt vid produktion av text, kommer först och följs av de friare exemplen som vanligen utgörs av hela satser. Idiomen som är besläktade med kollokationerna behandlas för sig som sublemman. På detta ges tydliga exempel i inledningen (SO:XVIIf.) Behandlingen av kollokationerna för sig är till fördel eftersom det, om dessa ingår i exempelsatser, kan vara svårt att avgöra kollokationernas räckvidd (Malmgren 2009:98f.). Rent typografiskt vore en tydligare gräns önskvärd å ena sidan mellan morfologiska och syntaktiska exempel och å den andra mellan kollokationerna och de fria illustrationerna. Som det nu är, är det svårt att skilja de olika typerna av exempel åt eftersom både de enskilda exemplen och de olika typerna skiljs åt av semikolon.

Också en rutinerad ordboksanvändare tvingas läsa de morfologiska exemplen för att småningom komma fram till exempelvis kollokationen *ingå/träffa avtal* eller den friare illustrationen till uppslagsordets användning: *~et mellan Sveriges Radio och staten*. En om möjligt tydligare gränsdragning av något slag hade sålunda varit att föredra. I inledningen kommenteras uppställningen (SO:XVII) men det är ju ingen hemlighet att denna sällan läses av ordboksanvändarna.

De språkliga exempel som ges är enligt SO ofta citerade, ibland något modifierade (SO:XVII) men egentligen är typen av mindre betydelse, eftersom huvudsaken är att exemplen fungerar som avsett. Det kan till och med hävdas att de modifierade exemplen är mera autentiska än de som tagits direkt ur korpora eftersom dessa är ryckta ur sitt sammanhang medan de modifierade befinner sig i den kontext de skapats för. (Nikula, H. 1995:311, Atkins & Rundell 2008:457). Språkproven kan ha semantisk, syntagmatisk, konnotationell och encyklopedisk funktion. (Malmgren 1994:109–111, Nikula, H. 1995:312f., Atkins & Rundell 2008:454f.). I normalfallet illustrerar språkproven flera än en funktion på samma gång.

Exemplen ska för att vara verkningsfulla vara naturliga och typiska, de ska vara informativa och de ska vara begripliga (Nikula, H. 1995:314–316, Atkins & Rundell 2008:458–461). Över lag är exemplen i SO informativa och bidrar både till förståelsen av uppslagsordet och visar detta i användning, men ibland skulle man önska sig vissa tillägg. *Demonstrera* v. 'förevisa och samtidigt förklara funktionssätt o.d.' följs av exemplen *han ~de den nya kopiatorn.* Av exemplet framgår att *demonstrera* är ett transitivt verb men med tanke på prepositionsanvändningen kunde till exemplet ha fogats *för studenterna*. I detta och många andra fall kompenseras likväl denna och liknande brister av valensuppgifterna där prepositionen finns angiven. Som förklaring till **C-uppsats** ges 'akademisk uppsats på C-nivå' och användningen förtydligas med fraserna *skriva ~, försvara sin ~*. Betydelsebeskrivningen är kort

och för den som inte är insatt i universitetsstudier knappast tillräcklig med tanke på produktion av text. I detta och liknande fall är det därför viktigt att SO konsekvent anför och förklarar det/de ord som kan förväntas skapa problem, i detta fall **C-nivå** 'kunskapsnivå motsvarande den tredje terminens studier i ett universitetsämne'. De exempel som knyter an till den kulturella kontexten fyller oftast de flesta krav som ställs på ett exempel på samma gång: de är informativa, naturliga och lätta att begripa. Bland exemplen i anslutning till **häxa** s. finns bl.a. *~n i pepparkakshuset* och *häxorna troddes fara till Blåkulla på sina kvastar*, användningen av **räv** illustreras med bl.a. *Mickel ~ lurade alla hundarna* och **dopp i grytan** med [...] *en traditionell jul med skinka, lutfisk och dopp i grytan*.

Ur genussynvinkel kan konstateras att en del av den gamla surdegen rensats bort bland exemplen. Både i SOB och i NEO fanns det gott om exempel där kvinnan sågs som beroende av mannen om hon över huvud tilläts visa sig (Nikula, K. 1997) men numera är kvinna och man mer jämspelta (ex. *han beter sig som ett svin mot sin fru* [...] resp. *hon var full som ett svin*). Förbättringen i SO, där också kvinnor kan göra saker på egen hand, bidrar knappast i någon större utsträckning till ordbokens användbarhet men bidrar inte heller till spridande av stereotypa uppfattningar om könen.

6. Idiom

I en enspråkig ordbok som ska gå att använda för reception av text är idiom ett självskrivet inslag – produktion av text låter sig göra även utan att idiom används. I enspråkiga ordböcker som är avsedda för både reception och produktion måste likväl åtminstone de vanligaste idiomen anföras (Svensén 2004:245); också modersmålstalare kan behöva försäkra sig om att ett idiom förekommer i den form och med den betydelse som man föreställer sig. I de flesta

ordböcker förekommer idiom i viss utsträckning men är ofta svåra att hitta. SO:s behandling av idiomen är föredömlig. De är lätta att hitta eftersom de införts som sublemman under sitt huvudord med fet stil men i något mindre grad än vanliga uppslagsord. Detta är naturligt eftersom idiomen utgör självständiga enheter med en betydelse som inte kan härledas som summan av de ingående ordens betydelser (idiom kan ofta också ges bokstavlig tolkning, t.ex. *hugga i sten*). Idiomen följs dessutom både av betydelsebeskrivning och av exempelsatser. I förtexten ges därtill enkla regler för placeringen: ett idiom behandlas under det första substantiv som ingår i detta utom i fall substantivet står i genitiv: jfr (**det är inte lätt att**) **lära gamla hundar sitta – vara på** (**sin**) **mammas gata** se **gata**. Saknas substantiv ska idiomet gå att hitta i artikeln om det första verbet, ex. **sopa hem ngt** se [1]**sopa**. Idiom som påminner om en liknelse med ett adjektiv som huvudord behandlas under adjektivet, ex. **gammal som gatan,** men finns också upptaget under **gata** med hänvisning till **gammal 1**. Vissa idiom kan vara svårare än andra att hitta, ex. **gå som smort**. Den som börjar sin sökning med **smort** hänvisas därifrån till **smörja** där idiomet behandlas. Även om man inte känner till reglerna för placeringen torde det ändå gå att hitta rätt eftersom idiomen, utan kommentar men med hänvisning, också ofta finns uppförda under andra lemman än där de enligt reglerna hör hemma och behandlas.

Trots att idiomen allmänt beskrivs som fasta fraser kan de uppträda i varierande former och variera antingen systematiskt eller kreativt (Szczepaniak 2006:36). Systematisk variation, som bland annat kan innebära användningen av synonymer eller strykningen av ett ord, noteras i viss utsträckning i SO, ex. **hamna/komma i orätta händer, komma** (**väl**) **till pass**. De kreativa varianterna, som man i motsats till de föregående skapat medvetet för att uppnå en viss, vanligen humoristisk effekt kan vara högst tillfälliga och noteras till följd av detta inte i ordböcker (Szczepaniak 2006:35–38) och sålunda inte heller i SO (SO:XVIII).

7. Konstruktionssätt

SO innehåller också formaliserade uppgifter om valens för i första hand verb men i vissa fall också för adjektiv och substantiv. I Norden var NEO först med denna typ av information och informationen har förbättrats i SO. Valensuppgifterna består av uppgifter om vilka aktanter som kan slutas till ett givet lemma och i vilken mån dessa är obligatoriska eller fakultativa. Formuleringen av valensuppgifterna är emellertid inte alltid oproblematisk. Bland annat är gränsen mellan de obligatoriska och fakultativa, inre aktanterna vag. Därtill är också gränsen vag mellan de fakultativa och de s.k. yttre aktanterna eller adverbial av olika slag som inte ska anges i konstruktionsordböcker (Malmgren & Toporowska Gronostaj 2009:184). Jag har bland annat därför inte i någon större utsträckning granskat i vad mån synonyma eller nästan synonyma grupper av verb uppvisar samma konstruktionsmönster. I stället har jag valt att närmare studera om man i SO lyckats presentera valensen så att uppgifterna är begripliga för folk i allmänhet trots att, som Bergenholtz & Vrang (2005:175f.) konstaterar i sin recension av Den Danske Ordbog, man kan utgå ifrån att denna information främst intresserar speciella grupper som ”akademikere og sprogliebhavere”.

Verbet *köpa* brukar användas som något av ett paradexempel då det gäller beskrivningen av valens. Betydelsen av ***köpa*** anges i SO som ’skaffa sig (ngt) som sin egendom mot ersättning’, dvs. SO utgår vid betydelsebeskrivningen från att verbet är treställigt med ”köpare-vara-pris”. Konstruktionsuppgiften skrivs däremot: ~ *ngt* (*av ngn*) (*till* el. *åt ngn*) (*för* BELOPP), vilket innebär att i denna förekommer fem element, ”köpare-säljare-mottagare-vara-pris”. En användare som är van vid att handskas med ordböcker kan sannolikt räkna ut hur valensuppgiften ska tolkas. Mindre vana användare kan däremot stöta på problem även om parenteser

är ett vedertaget sätt att markera sådant som är optionellt. Med tanke på folk i allmänhet har man därför i Den Danske Ordbog gått in för få valensbundna aktanter, vanligen två och på sin höjd fyra (Lorentzen & Trap-Jensen 2005:255). Språkprov med alla fem element i samma sats skulle knappast vara till någon större nytta eftersom de skulle bli både långa och föga naturliga. Flera kortare exempel där de olika elementen i tur och ordning kommer till uttryck (jfr Svenskt språkbruk) kunde däremot förbättra begripligheten men låter sig av utrymmesskäl knappast göra i allmänna (pappers)ordböcker (Malmgren & Toporowska Gronostaj 2009:189). Motsvarande konstruktionsmönster som vid *köpa* återkommer vid motsatsen till **köpa**, dvs. **sälja**, som beskrivs: ~ *ngt* (*till ngn* el. *ngt*) (*för* BELOPP) och **köp** s. som beskrivs på följande vis: ~ (*av ngt*) (*åt* el. *till ngn*) (*för* BELOPP). I dessa fall är konstruktionsuppgifterna förhållandevis genomskinliga. Mera komplicerade är konstruktionsuppgifterna för **läsa** 'låta blicken löpa längs textrader för att tillägna sig deras innehåll'[4] med valensen ~ (*ngn* el. *ngt el. SATS*) (*för ngn*), ~ (*om ngn* el. *ngt* el *SATS*) (*för ngn*), ~ (*i ngt*). Ordningsföljden mellan de anförda aktanterna kan inte uppfattas som ordföljdsregler, eftersom ordföljden är beroende av många olika faktorer, men utgångspunkten i ordboken förefaller naturligt nog vara något slags "normalordföljd".

Valensbeskrivningar kan till följd av sin komplexitet skapa tolkningsproblem (t.ex. *läsa*). Ett problem av annan typ är övergeneralisering. Det är nödvändigt att skilja mellan animata (*ngn*) och icke-animata (*ngt*) aktanter men inte sällan övergeneraliseras *ngt*. Malmgren & Toporowska Gronostaj (2009:184) konkretiserar detta med valensangivelserna till *läsa* v.; man kan läsa *ngt* men inte vad som helst vilket enligt dem gör att *ngt* borde preciseras. Konstruktionsuppgiften vid **läsa**, ~ *ngn*, illustreras med *vår mest läste författare*, men ett exempel med direkt objekt, t.ex. *läsa Strindberg*, vore ur användarens synvinkel tydligare. Också SATS kan skapa

4 I detta fall framgår av betydelsebeskrivningen som Nina Martola påpekat i e-brev (sommaren 2010) vilken typ av objekt som är möjlig.

problem i valensuppgiften till *läsa* v., och överallt där SATS ingår, eftersom det inte framgår vilken typ av bisats som kan komma i fråga – en bisats inledd med, *att*, *hur*, *när* o.a.? Detta problemen kunde åtminstone delvis ha avlägsnats med hjälp av exempel. En kontroll av tio slumpmässigt valda verb (*bönfalla*, *fråga*, *gnälla*, *prisa*, *rådfråga*, *satsa*, *stå ut*, *tala, yttra sig, önska*), visar att det endast i tre av artiklarna finns exempel som innehåller bisatser. Vid t.ex. *önska* v. innehåller artikeln däremot tre exempel där bisatsen inleds med *att*. I fall som detta kunde man ha varit mera sparsam för att i stället anföra exempel där sådana saknas, (t.ex. *satsa* v.). Både i fråga om övergeneralisering av *ngt* och om SATS kunde exemplen ha valts så att användaren med deras hjälp kunde sluta sig till vad *ngt* kan stå för (vid t.ex. *sjunga* v. preciseras med exempel vad som är ”sångbart”) och vilka bisatser som är möjliga. Denna lösning är likväl utrymmeskrävande och därför kunde man åtminstone vid SATS kanske i framtiden överväga någon annan typ av notation, t.ex. att/hur/när-SATS e.d. utan att ordboken för den skull förvandlas till en grammatik (jfr Lorentzen & Trap-Jensen 2005:252). Men inte heller detta är problemfritt (se Malmgren & Toporowska Gronostaj 2009:189). Problemet ska likväl inte överdrivas för det visade sig i praktiken vara oväntat svårt att hitta valensuppgifter där SATS ingår. Om det förhåller sig som Malmgren & Toporowska Gronostaj (2009:186) antar att kanske knappt tio procent av verben har komplicerade valensangivelser kan man konstatera att informationen vid de flesta verb trots allt är relativt lättforcerad.

Valensuppgifterna ges i SO i slutet av artikeln och endast en gång per betydelsemoment. Här frågar man sig om det inte vore bättre att omedelbart efter betydelsen också ange valensen eftersom det då skulle vara möjligt att se valensuppgifterna konkretiserade i de efterföljande exemplen. Eftersom placeringen ändå är densamma som i NEO får man anta att den inte utsatts för kritik av ordboksanvändare och recensenter utan att den ansetts fung-

erat någorlunda väl. Trots att valensuppgifterna inte alltid motsvaras av exempel och valensuppgifterna i vissa fall kan vara svårforcerade, är jag likväl helt överens med Malmgren & Toporowska Gronostaj (2009:186) om, att detta inte hade varit en tillräcklig orsak att helt avstå från valensuppgifter. För det är ingalunda enbart lingvister och eventuellt språkbrukare med annat modersmål än svenska som kan ha nytta av valensuppgifterna utan också modersmålstalande som är måna om sitt språk kan tveka vid t.ex. valet av preposition.

8. Första belägg, etymologier och sentenser

Även om varken första belägg eller etymologi underlättar reception eller produktion av text breddar uppgifterna i SO användarens kunskap om det svenska språket – etymologi uppges därtill av Atkins & Rundell (2008:208) numera tillhöra den information som kan förväntas ingå i medelstora, enspråkiga ordböcker. Intresset för ords ursprung är dessutom stort bland folk i allmänhet och även en knapphändig uppgift kan vara till nytta och motarbeta uppkomsten av alltför fantasifulla folketymologier. Dessutom kan etymologiska uppgifter säkerligen i många fall ge en djupare förståelse för ordets nuvarande betydelse och användning.

Efter de etymologiska uppgifterna följer på sina ställen ett citat som är nära förknippat med uppslagsordet och som är avsett att ge detta extra belysning. Citaten härstammar ofta från den svenska litteraturens klassiker i vid mening: det är inte bara Strindberg som fått bidra utan här finns också citat från Evert Taube, Povel Ramel, Hasse och Tage och många andra. Citaten aktualiserar betydande delar av kulturarvet men man kan misstänka att det är de något äldre ordboksanvändarna som känner sig mest hemmastadda bland upphovsmännen till citaten. Detta förringar självfallet inte den information som citaten ger om uppslagsordet.

Man kan fråga sig om en ordbok kan vara underhållande? SO är det i alla fall inte minst tack vare just citaten. Många är säkert de som av denna anledning och som utan ett direkt behov av information kommer att bläddra i SO och läsa en artikel här och ett citat där. Också uppgifterna om första belägg och ordens etymologi bidrar, förutom med den konkreta nyttan, också med ett visst underhållningsvärde.

9. Språkvårdsrutor

Nytt i SO i jämförelse med NEO är de s.k. stilrutorna som mot skuggad bakgrund ställvis uppträder allra sist i en artikel. I en kommentar till SO konstateras att ordboken är deskriptiv och att man åtminstone inte, om två uttryckssätt är lika frekventa och användbara, anser sig ha befogenhet att verka normerande (Malmgren 2009:97). I rutorna, där klassiska språkvårdsfrågor i anslutning till uppslagsordet diskuteras, ges likväl ofta rekommendationer. Rekommendationerna kan i princip gälla vilket språkligt fenomen som helst, t.ex. uttalet som i fallet *anrika* v. Om detta sägs: ”Det inte ovanliga uttalet med långt a är inte korrekt. Ordet har ingenting med *anrik* att göra utan vi har att göra med samma *an* som i t.ex. *gå an*.” I detta exempel nöjer man sig likväl inte med att rekommendera utan får sägas peka med hela handen. I fråga om *beslut* konstateras: ”Substantivet *beslut* kan kombineras med två olika s.k. funktionsverb, *fatta* och *ta*. Uttrycket *ta beslut* har tidigare ansetts sämre än *fatta beslut* men numera får de båda uttrycken anses likvärdiga.”

För den som inte har omedelbar tillgång till handböcker är det mycket välkommet att återkommande problem getts en synlig plats och fått en lösning. Språkvårdsrutorna, som enligt uppgift är 400 till antalet, kan betraktas som en strävan att ge första hjälpen; någon kan kanske tycka att de helt kunde ha utelämnats till

förmån för flera uppslagsord eller exempel. Förutom att rutorna finns är det välkommet att man, där det är möjligt, klart säger ifrån vad som accepteras (jfr ovan). Detta besparar ordboksanvändaren konsultationer av annan litteratur åtminstone i de tämligen enkla fall det här är frågan om. Vissa rekommendationer får mig likväl att hoppa till. Till dessa hör *färst* som superlativform till *få* om vilket konstateras: "*Färst* är fullt brukbart som superlativform." *Färst* får mig att tänka på någon som inte lärt sig att tala rent – men det står mig ju dess bättre fritt att också i fortsättningen undvika detta. På det hela taget är stilrutorna välvalda och bidrar dessutom till att ordboken med tanke på layouten ger ett luftigt intryck.

10. Motsvarar SO sitt syfte?

SO uppges vara en ordbok för alla. Att man verkligen bemödat sig om att ta hänsyn till användarna framgår redan av den innehållsrika men samtidigt pedagogiska inledningen. I denna går SO användaren till mötes bl.a. genom att man avstått från termer som kan vara okända för folk i allmänhet. Uppgifter om uttalet ges inte heller med IPA, som gemene man inte kan förväntas behärska, utan med det normala alfabetet, och de avvikande tecken som använts finns uppförda på en särskild lista med uttalsbeteckningar.

Uppslagsorden svarar väl mot det som användarna kan tänkas behöva i en ordbok för allmänt, icke specialiserat, bruk och man har skyggat varken för högt eller lågt. Till följd av denna stora spännvidd kan SO också spela roll som dokumentationsordbok. Tack vare att partikelverb ofta uppförts som lemman och att idiom är uppförda som sublemman (jfr avsnitt 4) behöver användaren inte längre plöja genom massor av text för att småningom hitta det han söker. Betydelsebeskrivningarna är så vitt jag kunnat se korrekta och tillräckliga för reception av text. Det förekommer visserligen i ordboken definitioner som med naturnödvändighet

förutsätter specialkunskaper eller som annars kan vara svåra att förstå men i sådana fall bidrar dess bättre ofta de språkliga exemplen till att identifiera referenten.

Produktion av text förutsätter i praktiken så mycket och varierande information att det vore orealistiskt att avkräva en ”vanlig”, och till och med en inlärningsordbok, allt som behövs (Atkins & Rundell 2008:410f.). Icke desto mindre bör den information SO lämnar mycket väl förslå för produktion av text. I synnerhet för ovana skribenter är informationen om ortografi och böjning samt ordbildning viktig med tanke på produktion. Synonymer och uppgifter om stil och bruklighet samt inte minst valensuppgifter är därtill nödvändiga. Uppgifterna om valens kan till en början förefalla svårbegripliga men är i själva verket i de flesta fall rätt enkla och genomskinliga. Som textproducent är man ju inte heller hänvisad enbart till dessa utan får också med hjälp av språkexempel delvis eller fullständigt samma information som i de formaliserade konstruktionsuppgifterna.

SO är en synkron ordbok men inte heller den etymologiskt intresserade blir lottlös. Även om de etymologiska uppgifterna inte kan vara alltför omfattande bidrar de förhoppningsvis till att hos allmänheten förhindra uppkomsten av alltför vidlyftiga etymologier. Tillsammans med språkvårdsrutorna, som är en nyhet i SO, gör de SO till en i hög grad mångfunktionell ordbok.

Även om layouten inte är en ordboksfunktion är den av största betydelse, då den kan medverka till snabb tillgång till data och på detta sätt öka ordbokens värde som ett praktiskt instrument (Gouws 2009:12f.). Att texten är uppställd i två spalter och inte i tre gör också att SO i jämförelse med sin föregångare är lätt att navigera i.

11. Sammanfattning

Efter denna rätt omfattande granskning kan jag bara konstatera att SO är en förträfflig ordbok som täcker de flesta behov en ”normal” ordboksanvändare kan tänkas ha. Merparten av den information som erbjuds tar visserligen sikte på produktion men detta hindrar inte att recipientens behov också tillfredsställs. Att SO bitvis därtill kan vara underhållande gör den inte sämre. Karakteristiskt för SO är att den är generös med information och den är lätt att orientera sig i. Den är kort sagt en guldgruva.

Litteratur

Atkins, B.T. Sue & Michael Rundell 2008: *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bergenholtz, Henning 2005: Falsche und richtige lexikographische Definitionen. I: Gottlieb, Henrik, Jens Erik Mogensen & Arne Zettersten (red.): *Symposium on lexicography XI. Proceedings of the Eleventh International Symposium on Lexicography May 2–4, 2002 at the University of Copenhagen*. Lexicographica. Series Maior. Tübingen: Niemeyer Verlag, 125–132.

Bergenholtz, Henning & Vibeke Vrang 2005: Den Danske Ordbog bind 2 (E–H) og 3 (I-L) – en ordbog for folket eller for akademikere? I: *LexicoNordica* 12, 169–187.

Engelberg, Stefan & Lothar Lemnitzer 2009: *Lexikographie und Wörterbuchbenutzung*. 4. überarb. u. erw. Auflage. Tübingen: Stauffenberg Verlag.

Gouws, Rufus H. 2008: Sinous lemma files in printed dictionaries: Access and lexicographic functions. I: Nielsen, Sandro & Sven Tarp (red.): *Lexicography in the 21st Century. In honour of Henning Bergenholtz*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 3–21.

Lorentzen, Henrik & Sanni Nimb 2010: Fra ordbog til wordnet. Hvordan udmøntes en traditionel ordbogsdefinition i en formaliseret wordnetbeskrivelse? I: Lönnroth, Harry & Kristina Nikula (red.): *Nordiska studier i lexikografi 10*. Tammerfors: Skrifter utgivna av Nordiska föreningen för lexikografi 11, 329–344.

Lorentzen, Henrik & Lars Trap-Jensen 2005: Grammatiske oplysninger i *Den Danske Ordbog*. I: Fjeld, Ruth Vatvedt & Dagfinn Worren (red.): *Nordiske studiar i leksikografi 7*. Oslo: Skrifter utgivna av Nordiska föreningen för lexikografi 8, 252–266.

Lundbladh, Carl-Erik 1999: Ordbok eller encyklopedi – en fråga om hänsyn till användaren? I: Slotte, Peter, Pia Westerberg & Eva Orava (red.) *Nordiska studier i lexikografi 4*. Helsingfors: Skrifter utgivna av Nordiska föreningen för lexikografi 5, 265–273.

Malmgren, Sven-Göran 1994: Språkprovens form och funktion i svenska betydelseordböcker från Östergrens Nusvensk ordbok till Svensk ordbok. I: *LexicoNordica* 1, 107–117.

Malmgren, Sven-Göran 2009: On production-oriented information in Swedish monolingual defining dictionaries. I: Nielsen, Sandro & Sven Tarp (red.): *Lexicography in the 21st Century. In honour of Henning Bergenholtz*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 93–102.

Malmgren, Sven-Göran & Maria Toporowska Gronostaj 2009: Valensbeskrivning i svenska ordböcker – och några andra. I: *LexicoNordica* 16, 181–196.

Martola, Nina 2010: Svensk ordbok utgiven av Svenska Akademien. 2009. 3732 s. ISBN 978-91-302267-3. I: *Folkmålsstudier* 48 (2010), 191–202.

Nikula, Henrik 1995: Exemplens funktion i ordböcker. I: Svavarsdóttir, Asta, Guðrún Kvaran & Jón Hilmar Jónson (red.): *Nordiska studier i lexikografi* 3. Reykjavík: Skrifter utgitt av nordisk forening for leksikografi. Nr 3, 311–320.

Nikula, Kristina 1997: Språkprov i könsperspektiv. I: *Fackspråk och översättningsteori*. VAKKI-symposium XVII. Vörå 22–23.2.1997. Vasa: Vasa universitet, 194–206.

NEO = *Nationalencyklopedins ordbok* 1995–1996. Höganäs: Bra Böcker.

SOB = *Svensk ordbok* 1986. Stockholm: Esselte Studium.

Sundman, Marketta 2010: Prepositioner som partiklar – ett grammatiskt kontinuum i svenskan? I: Byrman, Gunilla, Anna Gustafsson & Henrik Rahm (red.) *Svensson och svenskan. Med sinnen känsliga för språk*. Festskrift till Jan Svensson den 24 januari 2010. Lund, 309–318.

Svenska Dagbladet 9.12.2009 = *Borg ökar trycket på grekerna*. (http://www.svd.se/naringsliv/nyheter/borg-okar-trycket-pa-grekerna_3913629.svd). Hämtat 7.3.2010.

Svensén, Bo 2004: *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. Stockholm: Norstedts Akademiska Förlag.

Szczepaniak, Renata 2006: *The Role of Dictionary Use in the Comprehension of Idiom Variants*. Lexicographica. Series Maior. Tübingen: Max Niemeyer.

Kristina Nikula
professor em.
Puolalaparken 1 b B 12
FI-20100 Åbo
f2krni@uta.fi

Finn Stefánsson: Symbolleksikon

Loránd-Levente Pálfi

Stefánsson, Finn: *Symbolleksikon*. Grafisk tilrettelægning: Susanne Korsager. Forlagsredaktion: Charlotte Glahn. Billedredaktion: Jeppe Branner. 1. udgave, 1. oplag. København: Gyldendal 2009. 608 sider. Illustreret. Med hård indbinding. DKK 399,95.

1. Indledning: danske symbolleksika fra nyere og nyeste tid

Foruden det ovenfor anførte *Symbolleksikon* findes der følgende leksikografiske opslagsværker i helfagskategorien[1] fra nyere og nyeste tid på dansk om symboler: *Gads Symbolleksikon* (Biedermann 1991), *Symbolordbog* (Cirlot 2002), *Symboler og ideogrammer* (Liungman 2003) og *Politikens Symbolleksikon* (Cooper 2006). Biedermanns værk er forholdsvis lille af omfang (behandler 300 symboler); det er oversat fra tysk. Cirlots ordbog har et mysticistisk udgangspunkt og fokus og distancerer sig derfor fra almindelig akademia; den danske udgave er oversat fra spansk og i øvrigt produceret i den kirkelige højrefløjs regi. Liungmanns bog er forholdsvis omfattende (2.000 tegn/ideogrammer fra den vestlige verden), men omhandler kun ideogrammer og ikke symboler; den

1 Afhængigt af om en fagordbog – eller et fagleksikon – behandler et helt fagområde (fx maskinteknik), en del af et fagområde (fx pumpeteknik) eller flere fagområder (fx teknik), kaldes den helfagsordbog (eller enkeltfagsordbog), delfagsordbog eller flerfagsordbog (hhv. helfags-/enkeltfagsleksikon, delfagsleksikon eller flerfagsleksikon) og tilhører følgelig helfagskategorien, delfagskategorien eller flerfagskategorien af fagordbøger/fagleksika.

er oversat fra svensk. Coopers leksikon har fokus på symboler fra især kunstens verden (1.500 symboler); det er oversat fra engelsk.

I delfagskategorien findes *Ægyptiske guder og symboler* (Lurker 1997), som er oversat dels efter den tyske originaludgave og dels efter den engelske oversættelse, samt *Motiver og symboler i europæisk kunst* (Funder 2004). I flerfagskategorien findes *Hvorfor det? Talemåder & traditioner, skikke & symboler* (Eilertsen/Eilertsen 2005), som er en populærvidenskabelig udgivelse.

Med undtagelse af Cirlots leksikon, som kun er sparsomt illustreret (50–60 illustrationer fordelt over værkets 528 sider), er alle de ovennævnte værker rigt illustrerede og videnskabeligt funderede (omend i forskellig grad). Funder (2004), Liungman (2003) og Lurker (1997) kan bruges af både læg- og fagfolk. De øvrige af værkerne er primært til brug for lægfolk.

Det er unuanceret og på sin vis uforeneligt med den leksikografiske funktionslære at mene, at det, der p.t. foreligger på dansk, enten er for gammelt (Biedermann 1991), for småt (Biedermann 1991), med et for specifikt udgangspunkt eller en for snæver interesse (Cirlot 2002; Cooper 2006), med fokus på et bestemt genstandsområde (Liungman 2003; Lurker 1997) eller ikke tilstrækkeligt fagligt funderet (Cirlot 2002; Eilertsen/Eilertsen 2005). Til gengæld er det hævet over enhver diskussion, at Stefánssons *Symbolleksikon* er den overhovedet eneste danske originalproduktion i helfagskategorien i hvert fald de seneste 20 år, dvs. det eneste værk, der ikke er oversat fra et fremmedsprog, men produceret *af* danskere *for* danskere.

2. Finn Stefánssons symbolleksikon

2.1. Leksikografiske komponenter og virkemidler

Stefánssons *Symbolleksikon*, herefter SL, er rigt illustreret og indeholder 1.400 artikler, af hvilke 300 er rene henvisningsartikler

(se forordet, s. 6). Foruden opslagsdelen (s. 9–585) indeholder SL et "Forord" (s. 6–7), et "Efterskrift" (s. 587–605) og en kort "Litteraturoversigt" (607–608). Det fremviser således en simpel rammestruktur.

En stor mangel er det, at der ikke findes nogen brugervejledning, som bl.a. kunne oplyse om den ikke-typografiske strukturindikator > (bruges som henvisningsmarkør både artikelinternt og artikelfinalt). Til gengæld er det leksikografisk hensigtsmæssigt, at de termer eller ord, der figurerer som lemmata i opslagsdelen, og som omtales i efterskriftet, i efterskriftet er fremhævet med fed type. SL benytter sig her af den samme funktionsrelaterede, integrerede byggedel, jf. Tarp (1998:128), in casu en komponent-ekstern henvisningsstruktur, som også findes i *Gyldendals leksikon om nordisk mytologi* (Stefánsson 2005): Der gives eksplicitte komponent-eksterne henvisninger – nemlig fra efterskrift til opslagsdel. Men det er problematisk, at brugeren ikke får noget at vide om det. Heller ikke forordet er meddelsomt desangående eller med andre oplysninger vedrørende bogens brug. En mangel er det endvidere, at der ikke findes noget register. Et hurtigt gennemsyn af nogle vilkårlige artikler lader ane, at mængden af data i SL er væsentligt større, end de 1.400 opslagsord indikerer. Der kunne i et register sagtens have figureret i hvert fald 1.000 indførsler[2] (emneord, stikord og evt. egennavne, som ikke figurerer som opslagsord i opslagsdelen). Et simpelt register med 1.000 indførsler havde fyldt ca. 15 sider i SL's sidestørrelse; et mere sofistikeret, komplekst register havde fyldt omtrent 25 sider eller højst 35. Om vigtigheden af registre i leksikografiske opslagsværker se evt. Wiegand (2008). Som et konkret eksempel på problemet kan nævnes artiklen **skrift**: Her oplyses bl.a. om arabisk skrift, hieroglyffer, japansk skrift, kileskrift, kinesisk skrift, linear A, linear B og runer. Af disse er det

2 Tallet 1.000 kommer af følgende simple udregning: De fleste artikler indeholder mindst én behandlingsenhed, som ikke er identisk med den, som opslagsordet betegner.

imidlertid kun hieroglyffer og runer, der også figurerer i lemmaposition (førstnævnte som henvisningslemma, sidstnævnte som almindeligt lemma med egen artikel og med artikelfinal henvisning til **skrift**). Brugeren har ingen umiddelbar tilgang til de data, SL indeholder om arabisk skrift, japansk skrift, kileskrift, kinesisk skrift, linear A og linear B. Et register kunne have aflastet søgeruten; det havde været optimerende for tilgangsstrukturen og dataadgangsstrukturen. Et positivt forhold i denne sammenhæng er dog de mange henvisningslemmata. En mangel er det tillige, at der ikke findes nogen bibliografi (litteraturoversigten på s. 607–608 er ikke nogen bibliografi, men en kort vejledning til videre læsning). Dog bringes ved udgangen af nogle artikler en meget kort litteraturliste (se fx artiklen **sprog**).

2.2. Indhold (omtekster og opslagsdel)

I forordet hedder det, at "SL adskiller sig fra rækken af opslagsværker om symboler på det internationale marked – hvoraf nogle er oversat til dansk – på en række punkter" (s. 6):

> *Symbolleksikon* medtager naturligvis alle de gængse hovedsymboler og belyser dem med deres vedtagne symbolik. Men bogen rummer også en lang række artikler, hvor det kulturhistoriske og litterære er i centrum, fx artikler om typer (Don Juan, Aladdin, Snehvide, boheme m.m.), om skrift og sprog og om sport. I det hele taget dækker titlen over mere end den gængse betydning af symboler. Der er tale om at belyse den værdi, der tillægges et fænomen eller et begreb, og de følelser der er i forbindelse dermed, når de får et billedsprogligt (metaforisk eller symbolsk) udtryk. Det er menneskets billeddannende evne, der er i centrum. [s. 6]

Ambitionerne har ikke manglet. I ovenstående uddrag meddeles de overordnede principper for SL's dataselektion. Endvidere er følgende relevant i denne sammenhæng:

> Synsvinklen er både global og lokal. Især hvad angår det mytologiske og religiøse stof er der en global dimension; men bogen er også dansk, fx i de fleste af de litterære og sproglige eksempler. Den danske vinkel ses dog også i de mytologiske begreber og figurer, der får selvstændige artikler; her er det det jødisk-kristne, det græsk-romerske og det nordiske stof, der dominerer, netop fordi det har haft så dyb indvirkning på dansk kultur. [s. 6–7]

Det er ikke blot nyttigt (eksempelvis pga. en anmelders eller en potentiel brugers nysgerrighed), men også funktionelt vigtigt med en sådan udredning. På baggrund af den kan en potentiel bruger hurtigt afgøre, om vedkommende overhovedet kan/skal anvende det givne produkt som værktøj til informationssøgning; en bibliotekar kan afgøre, om bogen skal indkøbes til samlingen osv. En grundigere udredning havde dog været ønskelig, når nu SL-forfatteren alligevel er inde på emnet – eksempelvis en egentlig begrundelse for den brede forståelse af genstandsområdet (symboler og symbolik i det hele taget). Efter endt læsning af afsnittet tænker den potentielle bruger unægtelig: Hvorfor? Hvad er grunden til, at SL forfægter en så bred forståelse af symbolbegrebet? Er der tale om et generelt problemfelt inden for symbolforskningen? Der er blandt forskere måske ikke enighed om, hvordan genstandsområdet skal afgrænses? I metaleksikografien er der som bekendt ikke enighed om, hvordan genstandsområdet skal afgrænses (hvor går grænsen mellem leksikon og ikke-leksikon, mellem ordbog og ikke-ordbog?). Hvis det samme eller noget lignende er tilfældet inden for symbolforskningen, burde SL have gjort mere ud af at oplyse om det og evt. have henvist til relevant faglitteratur.

Den essayistiske stil, som benyttes i opslagsdelen, forekommer ikke ideel i leksikonartikler. Et grundlæggende element i essayistik er den bevidst manglende systematik (jf. Rasmussen et al. (2005:103), hvori essayet beskrives som en genre bl.a. kendetegnet ved "en friere, mere uformel og subjektiv måde at behandle et emne på, end den akademiske afhandling giver mulighed for"). Dermed ikke være sagt, at den akademiske afhandlings stil er ideel i leksikonartikler. Men systematik er netop noget, som generelt prioriteres højt i leksikografien. Hos Bergenholtz/Tarp (1994:153) findes følgende standpunkt:

> Derimod kan man under ingen omstændigheder anbefale den praksis med stereotype, ensartet opbyggede forklaringer, som findes i mange ordbøger à la følgende eksempel fra en dansk juridisk ordbog: [...] En sådan opbygning af de faglige forklaringer giver en stereotyp stil, der grænser til det kedelige, hvorfor det æstetiske indtryk og læseværdigheden forringes.

Jeg er mere enig med Puuronen (1995:268):

> Det förefaller konstigt att man sätter det estetiska intrycket framom en logisk uppbyggnad. De exempel som tas upp i detta sammanhang tycks till sin yttre struktur motsvara terminologiska definitioner som tydligt visar det beskrivna begreppets överbegrepp. At dylika definitioner borde undvikas i fackordböker är ett påstående som inte kan vara genomtänkt.

Jeg vil desuden supplere Puuronens argument med følgende påstand: Det er de færreste ordbogs- og leksikonbrugere, der slår op i et leksikografisk opslagsværk for underholdningens eller for adspredelsens skyld, eller fordi de vil læse værket fra ende til an-

den, som var det en roman eller en fagbog. Ordbøger og leksika er informationssøgningsværktøjer, dvs. de bruges til at opfylde specifikke typer af punktuelle informationsbehov hos specifikke brugergrupper. Derfor kan man til en vis grad gå ud fra, at kun et mindretal af brugerne vil opleve det som decideret generende med en nok så monoton stil, data(re)præsentation og databehandling – om end de æstetiske dimensioner ikke skal undervurderes. Til gengæld vil brugerne med stor sandsynlighed være utilfredse, hvis de oplever, at de leksikografisk kodificerede data i det givne opslagsværk er misvisende, ukorrekte, usystematiske og/eller mangelfulde.

Systematik er ganske vist mange ting, og en vis systematik findes i ethvert leksikografisk opslagsværk; det er dokumenttypologisk betinget, ellers er der slet ikke tale om et leksikografisk opslagsværk. Den slags systematik, jeg her efterlyser, gælder udformning, mængde og præcision af data(angivelser) i artiklerne. Det optimale er, at brugeren finder den/de omtrentligt samme type(r) data udformet og ordnet på omtrent samme måde i de forskellige artikler i det pågældende værk. Essayistik er derfor i sagens natur uhensigtsmæssig, når det gælder ordbogs- og leksikonartikler. Som et eksempel kan fremdrages artiklen **skrift**, som bl.a. indeholder følgende afsnit:

> Inden for zenbuddhismen ophæver kunstneren med sine pludselige penselstrøg, de sorte tuschtegn på det hvide papir, forskellen på subjekt og objekt: han *er* skriften og viser i sit arbejde den oprindelige buddha-natur. [kursiveringen er SL's egen]

Det er typisk for den essayistiske stil at benytte poetiske virkemidler eller i det hele taget poetisk inspireret sprogbrug ("han *er* skriften") og at kommentere det ene og det andet emne *en passant*. Problemet er imidlertid, at brugeren ikke bliver klogere på hverken

zenbuddhisme eller buddhisme efter at have læst ovenstående afsnit. Han/hun får strengt taget kun følgende at vide: Kunstnerens "pludselige penselstrøg, de sorte tuschtegn på det hvide papir" ophæver "forskellen på subjekt og objekt". Er der her tale om data, ud fra hvilke brugeren kan udlede information, som kan blive til reel viden og afhjælpe et vidensproblem? Nej. I samme artikel (**skrift**) findes også følgende afsnit:

> Hieroglyffernes gåde blev først løst i 1822; bl.a. derfor kaldes de "hellige tegn", og teorierne frem til løsningen var vidtløftige; noget tilsvarende gælder mht. den såkaldte kileskrift [...].

Her er det dels et problem, at brugeren i det hele taget kun får meget lidt at vide om hieroglyffer, dels at det ikke oplyses, hvem det var, der løste "hieroglyffernes gåde". I samme artikel (**skrift**) findes dog også et andet kort afsnit, som – korthed og essayistisk stil til trods – er vellykket:

> I islam er Koranen Guds Ord, dvs. bogstaverne og det arabiske sprog er hellige. Da der er forbud mod at fremstille Gud og mennesker i islam, får skriften, det enkelte bogstavs ciselerede skønhed, en særlig manifestationskarakter af det guddommelige, fx i koranhåndskrifter og på vægge i moskeer og paladser.

Afsnittet er informativt og viser, at den essayistiske stil ikke ligefrem er helt uanvendelig i leksikonartikler. Men som det fremgår ovenfor, er der her ikke tale om smag og behag, men om noget så essentielt som leksikografisk data(re)præsentation og databehandling – dvs. aspekter, som har konsekvenser for værkets funktionalitet, hvilket her bl.a. betyder informationsudbyttet eller manglen på samme.

Et yderligere problemaspekt vedrørende indholdet er manglen på faglig tyngde. Det er imponerende, at én forfatter har skrevet hele værket, men det har konsekvenser, som ikke kan ignoreres: En lang række artikler havde utvivlsomt fremvist en større sikkerhed, dybde, detaljerigdom og præcision i emnebehandlingen, havde arbejdet været fordelt mellem flere fagfolk – fx fem, fire eller bare to skribenter inkl. Stefánsson selv. Der er principielt ingen grund til at have 50 eller 100 bidragydere til et leksikon, devisen "jo flere des bedre" holder ikke altid.

Emnevariationen i SL er mangfoldig, hvilket ikke alene hænger sammen med værkets genstandsområde, som er meget vidtspændende, men også med, at SL forfægter en ualmindeligt bred forståelse af symbolbegrebet. I forordet (s. 7) skriver forfatteren:

> Naturligvis er bogen primært objektivt informerende. Men den er også personlig. Jeg tillader mig at tolke og perspektivere og tage stilling. Bogen afspejler mine erfaringer og refleksioner: rejser kloden rundt, studier og læsning og undervisning inden for dansk (litteratur og sprog) og religion. Jeg har mine litterære favoritter og ideologiske idiosynkrasier, og det afspejles uden tvivl i bogen.

Det er rosværdigt og ikke mindst modigt med et sådant udsagn i et leksikon. De færreste leksikografer (eller videnskabsfolk i det hele taget) tør formulere noget sådant, selvom der i hvert fald inden for humaniora og samfundsvidenskaberne synes at herske en konsensus om, at reel objektivitet er umulig at efterleve. Alligevel er der en grænse: SL er ikke et decideret personligt leksikon, det fremgår med al tydelighed af indholdet.

2.3. Funktionalitet og brugergrupper

Ovenstående diskussion om essayistik (se kap. 2.2) bør ikke give

det indtryk, at SL gennemgående er informationsfattigt. Tværtimod er værket ganske informationsrigt. Et eksempel på en kort, udmærket og informativ artikel er **geomantik**:

> **geomantik** (græsk: jordspådom), den opfattelse, at jorden har bestemte kraft- eller energifelter, som svarer til kraften i universet som sådant eller til træk i folkets mytologi. Den spådomskyndige kan lægge og tyde særlige mønstre på jorden, fx håndfulde af sand. Geomantiske systemer kendes mange steder fra, fx hos yoruba i Vestafrika. Et særlig[t] udviklet system er det kinesiske >fengshui, der er knyttet sammen med opfattelsen af verden som et balanceforhold mellem yin og yang.

Artiklerne varierer en del i størrelse, men er som regel væsentligt længere end den her gengivne. SL er et kognitivt-funktionelt leksikon af den traditionelle slags, dvs. et leksikon med primærfunktionen "at yde hjælp til videnstilegnelse" (se bl.a. Tarp 2007). Artikelinitialt gives efter opslagsordet (i reglen kun i tilfælde af fremmedord) af og til semietymologiske data, fx angivelsen "(græsk: jordspådom)" i artiklen **geomantik** foroven. De optimerer værkets primærfunktion. At der netop gives semietymologiske data og ikke etymologiske data *sensu stricto* (fx "af gr. *gē* jord + *mant'eia* spådom" (Brüel/Nielsen 2003:216)), indikerer tydeligt hensyntagen til den læge bruger. Hvorfor de semietymologiske data ikke bringes konsekvent ved fremmedord (og heller ikke ved egennavne), er til gengæld uklart.

Om brugergruppe(r) meddeles intet i omteksterne. Om funktionalitet findes følgende bemærkning i forordet (s. 6):

> Bogen kan bruges i mange sammenhænge: give konkret information, til undervisning i især de litterære fag og religion og tværfagligt, være en bank af litterære eksempler, inspirere til videre læsning.

Den anvendte praksis med semietymologiske data sammenholdt med artiklernes sproglige udformning (letforståeligt sprog), indholdsmæssige tyngde (eller mangel på samme) og ikke mindst ordet "undervisning" i ovenstående uddrag indikerer, at værket primært henvender sig til lægfolk i almindelighed og gymnasieelever i særdeleshed (samt andre elever/unge på lignende skole-/videnstrin). Det kan sekundært også bruges af folkeskole- og gymnasielærere (samt andre lærere på tilsvarende niveau), såfremt disse i forbindelse med brugen af SL også konsulterer andre ressourcer. Forskere kan kun til dels benytte værket – såfremt det skal være i forskningsøjemed. SL er ikke noget videnskabeligt leksikon forstået som et leksikon udarbejdet *af* videnskabsfolk til brug *for* videnskabsfolk. Når forskere alligevel kan benytte det, omend kun til dels, skyldes det dets rigdom på henvisninger til skønlitteratur og dets aktualitet. Med henvisningerne til skønlitteratur tænkes selvsagt ikke på de sporadiske, artikelfinale bibliografiske referencer (typisk til faglitteratur), men på de mange – typisk skønlitterære – citater, der bringes inde i artiklerne (ofte indledningsvis), og som kan indeholde værdifulde oplysninger om, hvad den ene og anden fremtrædende tænker eller forfatter har skrevet om dette eller hint symbol. Med aktualiteten menes både det, at SL er udkommet i 2009, og at det bevidst har fokus på moderne tid og moderne fænomener. Fokus på det moderne hænger ikke sammen med udgivelsestidspunktet (SL kunne også have haft fokus på gammel tid, uagtet at værket er produceret i nyeste tid) – der er tale om et bevidst valg fra forfatterens side. Derfor findes der emnebehandlinger i SL, som er så aktuelle, at man ikke kan finde de oplysninger i et nok så videnskabeligt og nyt leksikon. Også den ualmindeligt

– for ikke at sige kontroversielt – brede forståelse af symbolbegrebet bevirker, at forskere med fordel kan søge informationer i SL, fordi det indeholder data, som med høj sandsynlighed ikke findes andre steder.

2.4. Form (opsætning og design generelt)

En meget positiv dimension ved SL er det grafiske og værkets design i det hele taget: Skrifttypen (Officina) er rolig og derfor behagelig for øjnene. Den sætter sig så stabilt på siden, at den løse bagkant ikke generer. Opsætningen er i det hele taget vellykket. Man kunne have ønsket en mindre luftig løsning angående sidernes indretning/opsætning, men en sådan er næppe mulig, når der opereres med så mange forskellige billedstørrelser, som tilfældet er i SL: Der bringes bl.a. helsides-, halvsides-, tredjedelsides- og fjerdedelsidesbilleder. Opsætningen er i det hele taget ambitiøs og innovativ. Opslagsordene er ikke rykket ud mod venstre i forhold til artiklerne, som det oftest ses i leksikografiske opslagsværker, men fremstår alligevel tydeligt pga. fremhævningen med fed type. Strukturindikatoren > forekommer elegant. Ulempen ved den er ganske vist, at brugeren ikke kan se den ved blot at *betragte* en given side – det er nødvendigt at *gennemlæse* en given artikel. For de artikelfinale henvisninger indledt med > er en gennemlæsning af den pågældende artikel naturligvis ikke nødvendig. Her er det dog et problem, at der ikke findes nogen brugervejledning, som oplyser om fx de artikelfinale henvisninger (se kap. 2.1 foroven).

Billeddimensionen er vellykket. Antallet af billeder er overvældende og udvalget mangfoldigt. Den hårde indbinding, som vel er nødvendig til en bog på denne størrelse, men af den grund ikke en selvfølge, er også prisværdig. Hvorvidt omslagets udseende er flot, kan diskuteres. Mange brugere vil formentlig være begejstrede over eksempelvis sølvfarven på vel at mærke hele bogblokken, andre vil mene, at den slags kun hører hjemme på salmebøger og

Biblen. Uanset hvad man personligt måtte mene om det, er det næppe diskutabelt, at designarbejdet i det hele taget (herunder også opsætning, billedudvalg osv.) er mere end almindeligt ambitiøst og alene af den grund rosværdigt.

3. Sammenfatning

SL er et anvendeligt værktøj for især unge på ungdomsuddannelser (herunder ikke mindst gymnasieelever) og lægfolk i almindelighed, om end også folkeskolelærere, gymnasielærere og andre lærere på tilsvarende niveau kan få glæde og nytte af det, såfremt de i forbindelse med brugen af værket også konsulterer andre kilder. Endelig kan forskere anvende det, men kun til dels: SL er ikke et videnskabeligt leksikon (dvs. ikke produceret *af* forskere til brug *for* forskere).

Af negative forhold ved SL bør fremhæves bl.a. (1) den manglende brugervejledning, (2) det manglende register og (3) den kendsgerning, at forfatteren har valgt at lave arbejdet alene (her ses bort fra grafikerens samt billed- og forlagsredaktørens indsats) i stedet for at deles om det med en større skribentkreds på fx fem eller bare to fagfolk.

Af positive forhold ved værket bør fremhæves (1) dets fokus på moderne tid, (2) dets ualmindeligt generøse billedudvalg samt (3) dets brede forståelse af symbolbegrebet. Sidstnævnte kan dog også være en ulempe afhængigt af den aktuelle bruger i den givne brugersituation: Brugeren kan eksempelvis blive forvirret over dataselektionen, hvis vedkommende er vant til at bruge traditionelle symbolleksika med en væsentligt snævrere symbolforståelse, og han/hun kan komme i tvivl om, hvad det egentlig er, man kan finde information om i SL. Men denne usikkerhed kan siges at gøre sig gældende for de fleste leksikografiske opslagsværker, hvis ikke alle: Brugeren kan aldrig med sikkerhed vide, *om* han/hun

kan finde den ønskede information i et givent værk, førend vedkommende har slået efter (og helst flere forskellige steder).

Layout- og i det hele taget designmæssigt er SL en flot, ambitiøs, men også dristig udgivelse.

Alt i alt er SL et brugbart værktøj og under alle omstændigheder en imponerende enmandspræstation.

Bibliografi

Leksikografiske opslagsværker

Biedermann, Hans 1991: *Gads Symbolleksikon*. København: Gad.

Brüel, Sven/Niels Åge Nielsen 2003: *Fremmedordbog*. 11. udgave, 9. oplag ved Lilian Plon. København: Gyldendal.

Cirlot, Juan-Eduardo 2002: *Symbolordbog*. København: Sankt Ansgar.

Cooper, Jean Campbell 2006: *Politikens Symbolleksikon*. 2. udgave, 5. oplag. København: Politiken.

Eilertsen, Mette/Mogens Eilertsen 2005: *Hvorfor det? Talemåder & traditioner, skikke & symboler*. København: Frydenlund.

Funder, Lise 2004: *Motiver og symboler i europæisk kunst*. København: Nyt Nordisk Forlag, Arnold Busck.

Hansen, Peter Nørbæk/Palle Qvist 2006: *Samfundslex*. 3. udgave, 1. oplag. København: Gyldendal.

Liungman, Carl G. 2003: *Symboler og ideogrammer*. København: Aschehoug.

Lurker, Manfred 1997: *Ægyptiske guder og symboler*. København: Politiken.

Pedersen, Mogens N./Kjell Goldmann/Øyvind Østerud (redaktører) 2004: *Leksikon i statskundskab*. København: Akademisk Forlag.

Rasmussen, Henrik (redaktør)/Kamilla Hygum Jakobsen/Jeanne Berman 2005: *Gads Litteraturleksikon*. København: Gad.

Stefánsson, Finn 2005: *Gyldendals leksikon om nordisk mytologi.* København: Gyldendal.

Anden litteratur

Bergenholtz, Henning/Sven Tarp (redaktører) 1994: *Manual i fagleksikografi. Udarbejdelse af fagordbøger – problemer og løsningsforslag*. Herning: Systime.

Pálfi, Loránd-Levente/Patrick Leroyer/Adam Wagner/Spiros Divaris Vesterdahl 2008: Skomager, bliv ved din læst! Om politologiske leksika, politik i leksika og leksikografiske værktøjer. I: *LexicoNordica* 15, 197–218.

Puuronen, Nina 1995: [Anmeldelse af] Henning Bergenholtz & Sven Tarp (redaktører): Manual i fagleksikografi. Udarbejdelse af fagordbøger – problemer og løsningsforslag. Herning: Systime 1994. I: *LexicoNordica* 2, 265–270.

Tarp, Sven 1998: Leksikografien på egne ben. Fordelingsstrukturer og byggedele i et brugerorienteret perspektiv. I: *Hermes : Journal of Linguistics* 21, 121–137.

Tarp, Sven 2007: Lexicography in the Information Age. I: *Lexikos* 17, 170–179.

Wiegand, Herbert Ernst 2008: Wörterbuchregister. Grundlagen einer Theorie der Register in modernen Printwörterbüchern. I: *Lexikos* 18, 256–302.

Loránd-Levente Pálfi
forskningsassistent
Center for Leksikografi
Handelshøjskolen, Aarhus Universitet
Fuglesangs Allé 4
DK-8210 Aarhus V
llp@asb.dk

KOMMENTARER
TIL TIDLIGERE BIDRAG

Nogle bemærkninger til Henning Bergenholtz: "Hurtig og sikker tilgang til informationer om ordforbindelser" i LexicoNordica 16

Christian Becker-Christensen

1. Indledning

LexicoNordica 16 havde som tema ordforbindelser i nordiske ordbøger. Blandt de behandlede emner var klassifikationen af ordforbindelser og præsentationen af de forskellige typer i diverse ordbøger.

Et af bidragene var Henning Bergenholtz' *Hurtig og sikker tilgang til informationer om ordforbindelser* (Bergenholtz 2009) der har brugertilgangen til ordforbindelser som emne. Dette specificeres nærmere sådan at "Tilgangen og ikke tilgangsstrukturen og heller ikke makrostrukturen er temaet" (op.cit.: 30). Artiklen omhandler en test (et "tilgangseksperiment") der registrerer en forsøgspersons "søgevej" og "tidsforbrug" ved søgning på 10 faste vendinger (idiomer og ordsprog) i otte forskellige danskordbøger hvoraf de fem er papirudgaver som angives at være:

Nudansk Ordbog (2005)
Den Danske Ordbog (2003-2005)
Ordbog over det danske Sprog (1918-1956)
Talemåder i dansk (1998)
Danske Talemåder (1998)

og de tre er elektroniske udgaver, nemlig:

netudgaven af *Ordbog over det danske Sprog* (2008)

og Henning Bergenholtz mfl.s to netordbøger:

Den Danske Netordbog (2008)
Betydning af faste vendinger (2009)

Opgaven der stilles, er at forsøgspersonen bliver bedt om "at finde ud af, hvad bestemte faste vendinger betyder, hhv. om probanden kan finde vendingen i udvalgte opslagsværker" (op.cit.: 34, 36). Alle de valgte ordbøger er blevet testet i formentlig en bestemt rækkefølge; det indgår altså ikke i testen at forsøgspersonen har skullet vælge hvilken ordbog der ville være mest relevant for besvarelsen af opgaven.

Som resultater af undersøgelsen kan udledes:

- at søgetiden er mindre når søgningen på en ordforbindelse giver positivt resultat end når den giver negativt resultat (jf. op.cit.: 44, 49),
- at elektroniske versioner (med direkte søgning på en ordforbindelse eller en del af den) er hurtigere at søge i end papirudgaver (og den elektroniske version af *Ordbog over det danske Sprog* hvor man skal scrolle gennem den fundne artikel) (jf. op.cit.: 48, 51),
- og at de involverede ordbøger fordeler sig i en rangordning efter "positivt resultat + tidsforbrug i gennemsnit" hvor *Betydning af faste vendinger* og *Den Danske Netordbog* får den bedste score, mens *Nudansk Ordbog* og *Talemåder i dansk* ligger i bunden (op.cit.: 50).

Bergenholtz' testbeskrivelse er gengivet i Hermes 44, 2010, i artiklen Bergenholtz & Gouws: "A New Perspective on the Access Process".

Jeg har disse bemærkninger til testens resultater og tilrettelæggelse:

2. Søgning med negativt resultat

Undersøgelsen omfatter opgørelse af forsøgspersonens tidsforbrug ved søgninger med såvel positivt som negativt resultat, jf. definitionen af søgetid som "tidspunktet, hvor behovet for en ordbogskonsultation opstår, hhv. når eksperimentets indhold fremlægges, indtil det herudfra opståede spørgsmål er blevet besvaret, hhv. at søgningen opgives resultatløs" (op.cit.: 33). Men i det sidste tilfælde, søgning med "negativt resultat", inkluderer undersøgelsen også søgning efter ordforbindelser som ikke forekommer i undersøgelsesmaterialet.

Fx bruger forsøgspersonen 2 minutter og 17 sekunder minutter på at lede efter vendingen "barn på gule plader" som ikke figurerer i *Nudansk Ordbog*, og 4 minutter og 36 sekunder på at lede forgæves i *Ordbog over det danske Sprog* hvor vendingen naturligt nok heller ikke forekommer. I de to specialordbøger *Danske Talemåder* og *Talemåder i dansk* som eksplicit ikke medtager ordsprog, giver søgning på "liden tue kan vælte stort læs" og "hvor godtfolk er, kommer godtfolk til" ligeledes lange søgetider.

I *Betydning af faste vendinger*, hvor de nævnte ordforbindelser findes, er søgetiden henholdsvis 15 sekunder, 16 sekunder og 16 sekunder.

Der redegøres ikke i undersøgelsen for i hvilke tilfælde de søgte sekvenser forekommer i undersøgelsesmaterialet, og i hvilke tilfælde de ikke eksisterer i materialet. Det er på denne baggrund uklart hvad der menes med "tilgang til information om ordforbindelser". Har "tilgang" at gøre med strukturen i ordbøgerne, altså på hvilken måde genstanden for en søgning præsenteres i ordbogen? Eller betyder "tilgang" at undersøgelsen drejer sig om ordbøgernes repræsentativitet: hvornår ordbogen er udkommet, dens størrelse,

hvilken type der er tale om, og hvilken brugergruppe den er tiltænkt?

Hvis man mener at en test af søgetider giver bedst mening i forhold til det første, bedømmelsen af og deraf videreudviklingen af ordbøgers tilgangsstruktur, så bør en test af søgetider være begrænset til sekvenser som findes i de testede ordbøger og bør ikke sammenblandes med en undersøgelse af om en bestemt sekvens forekommer i en bestemt ordbog eller ej.

Et andet forhold er at der med "tilgang til information om ordforbindelser" måles på om en ordforbindelse har en betydningsforklaring og ikke blot om man kan finde den pågældende ordforbindelse, dvs. fundne ordforbindelser uden betydningsforklaringer tæller negativt. Hvis man igen – trods Bergenholtz' indledende bemærkning om at temaet er "Tilgangen og ikke tilgangsstrukturen" (op.cit.: 30) – tænker sig at testen skulle have noget at gøre med tilgangen til ordforbindelser, så burde det alene være forekomst af sekvensen der var relevant for søgningen og ikke om der var angivelser knyttet til den.

Man kan venligt bemærke at disse forhold udgør metodiske uklarheder. Eller man kan spørge hvad man egentlig skal bruge testen til. Det er under alle omstændigheder vanskeligt at se værdien af den samlede opgørelse af gennemsnitssøgetider hvori indgår negativt resultat: "tidsforbrug i gennemsnit" og "negativt resultat + tidsforbrug i gennemsnit" (tabellen op.cit.: 50).

3. Udvalget af testede ordbøger

Ordbog over det danske Sprog er en historisk ordbog i 28 bind som spænder over perioden fra ca. 1700 til redaktionstidspunktet (1918-1956) med tilsvarende lange og layoutmæssigt komprimerede artikler; *Betydning af faste vendinger* er en elektronisk ordbog over netop ordforbindelser som dem der testes for. Sammenstillingen

og sammenligningen af så forskellige ordbogstyper i en og samme tilgangsundersøgelse er besynderlig. For at en test af søgetider skal have nogen mening, bør undersøgelsen forholde sig til typer af ordbøger ved at holde sig til en afgrænset type eller ved eksplicit at sammenholde forskellige typer og skal ikke sammenblande forskellige ordbogstyper på den måde som det gøres i testen. Inden for testens rammer kunne testmaterialet lige så vel have medinddraget Leths *Dansk Glossarium* (1800), som nævnes som eksemplet på en specialordbog til brug for tekstreception (op.cit.: 31).

4. Angivet og faktisk testede ordbøger

To ordforbindelser "Liden tue kan vælte stort læs" og "Hvor godtfolk er, kommer godtfolk til" giver negativt resultat for *Nudansk Ordbog* (med deraf følgende lange søgetider, henholdsvis 1 minut og 26 sekunder og 2 minutter og 12 sekunder) (op.cit.: 44-45).

Men begge står i den udgave af *Nudansk Ordbog* som ifølge præsentationen af forsøgsmaterialet (op.cit.: 35) og ifølge litteraturlisten (op. cit.: 53) er genstand for testen, nemlig 19. udgave, 2005:

tue ['*tu·ə*] sb. *-n, -r, -rne*
□ en lille forhøjning i jordoverfladen ◊ *engen var fuld af tuer · de måtte springe fra tue til tue for ikke at få våde fødder* ◊ ***græstue*** · ***myretue***
ORDSPROG: **liden tue kan vælte stort læs** ◊ *sagen kan blive den tue der vælter organisationens læs*

godtfolk sb.
□ (ældre) jævne, hæderlige mennesker ◊ *til stede var læger, sygeplejersker og andet godtfolk*

• **være kommet af godtfolk** (ældre) være af god familie
ORDSPROG: **hvor godtfolk er, kommer godtfolk til**

Nudansk Ordbog, 19. udgave, 2005

I stedet for den angivne 19. udgave er forsøgspersonen blevet forelagt en tidligere udgave af Nudansk Ordbog hvori ikke er medtaget informationstypen ordsprog, og hvor de to ordforbindelser derfor ikke forekommer. At der i testen ikke er tale om den angivne 19. udgave af *Nudansk Ordbog* fremgår desuden af side 44 hvor det siges at "probanden ikke kunne forstå, at der stod en betydning 1, men ikke nogen betydning 2". Denne praksis for markering af overgangen mellem hoved og krop med tallet "1" uanset om der fulgte flere betydninger efter eller ej, blev benyttet i Nudansk Ordbog 17. udgave, 1999, og 18. udgave, 2001. I Nudansk Ordbog 19. udgave markeres overgangen ved en firkant når artiklen ikke er opdelt i talnummererede betydninger.

Ud fra forsøgets præmis, at fundet af betydningen af ordforbindelserne er kriterium for måling af tidsforbruget ved søgning på ordforbindelserne, vil rettelse af denne fejl i testen dog ikke ændre på resultatet "negativ" i de to tilfælde.

5. Rangordning mellem ordbøgerne

Bergenholtz' sammenfattende tabel viser "en rangordning mellem ordbøgerne, som har et positivt resultat som vigtigste og søgetiden som næstvigtigste argument" hvilket "giver en rangordning, hvor ordbøgerne er anført med den – i denne test – bedste ordbog øverst" (op.cit.: 49-50). Her lander *Betydning af faste vendinger* som den bedste ordbog med 10 positive søgninger og et gennemsnit på 24 sekunder. *Nudansk Ordbog* placeres på en syvendeplads

med 5 positive søgninger og en gennemsnitssøgetid på 36 sekunder lige over *Talemåder i dansk* som har 3 positive søgninger med et gennemsnit på 30 sekunder.

Men hvis i stedet søgetid med positivt resultat, dvs. søgetid på de faktisk forekommende ordforbindelser i ordbøgerne, blev lagt til grund for en lidt mere meningsfuld opgørelse, ville ordningen efter gennemsnitlige søgetider være denne:

Betydning af faste vendinger	20”
Den Danske Netordbog	24”
Talemåder i dansk	30”
Nudansk Ordbog	36”
Den Danske Ordbog	40”
Ordbog over det danske Sprog (netudgave)[1]	41”
Danske Talemåder	48”
Ordbog over det danske Sprog (papirudgave)	2’ 35”

Bergenholtz viderebringer resultatet af sin test i den efterfølgende artikel i LexicoNordica 16: Bergenholtz & Bjærge (2009:59) med denne præsentation: “resultaterne af et eksperiment med søgninger i otte danske ordbøger efter 10 faste vendinger”. Men nu citeres testen alene med en rangordning der viser gennemsnitlige søgetider der (ukommenteret) inkluderer søgninger med negativt resultat, altså søgetider for søgninger på ikkeforekommende ordforbindelser. I rangordningen her står *Nudansk Ordbog* på en sjetteplads med en gennemsnitssøgetid på 1 minut og 3 sekunder, et resultat der er baseret på en metodisk dubiøs inkludering af resultater af søgning på ikkeeksisterende forekomster i testmaterialet.

1 At netudgaven af *Ordbog over det danske Sprog* har en så markant bedre søgetid end papirudgaven, skal nok ses i lyset af at forsøgspersonen har fået forelagt papirudgaven først, før versionen som netudgave, og dermed har fået kendskab til hvor i ordbogen de testede ordforbindelser findes (jf. rækkefølgen i eksperimentbeskrivelsen op.cit.:37 og i tabellerne for søgning på de enkelte ordforbindelser).

6. Afsluttende bemærkninger

Bergenholtz slutter sin artikel med at der i bidragene i LexicoNordicas temadel "er for meget fokus på selektering, da faste vendinger ikke kan selekteres fuldstændigt, hvis man da ikke kommer op på hundrede tusinde lemmatiserede eller kodificerede faste vendinger i den enkelte ordbog. I stedet bør der lægges større vægt på at forske i tilgang, særligt i søgetider, og hvordan en ordbogs layout kan bidrage til lavere søgetid, så der kan udarbejdes mere hensigtsmæssige ordbøger med en hurtigere og sikrere tilgang" (Bergenholtz 2009:51).

Men Bergenholtz' løst tilrettelagte tilgangseksperiment i LexicoNordica 16 giver netop ikke afsæt til dette her erklærede formål da han ikke (ud over sporadiske bemærkninger) forholder sig til tilgangsstrukturen i de testede ordbøger (fx placering, ordning og markering af ordforbindelserne) og søgetiderne derfor ikke kan relateres til forskelle i ordbøgernes layout.

Referencer

Bergenholtz, Henning 2009: Hurtig og sikker tilgang til informationer. I: *LexicoNordica 16*, 29-54.

Bergenholtz, Henning & Esben Bjærge 2009: Konception af fire monofunktionelle ordbøger med faste vendinger. I: *LexicoNordica 16*, 55-73.

Bergenholtz, Henning & Rufus Gouws 2010: A New Perspective on the Access Process. I: *Hermes – Journal of Language and Communication Studies no 44-2010*.

Betydning af faste vendinger 2009. www.idiomordbogen.dk.

Danske Talemåder 1998 = Allan Røder: *Danske Talemåder*. København: Gad.

Den Danske Netordbog 2009. www.ordbogen.com.
Den Danske Ordbog 2003-2005. København: Det Danske Sprog- og Litteraturselskab & Gyldendal.
Leth, Jens Høier 1800: *Dansk Glossarium*. Kiøbenhavn: Trykt paa Hofboghandler Simon Poulsens Forlag.
Nudansk Ordbog 1999 = *Politikens Nudansk Ordbog*. 17. udg. København: Politikens Forlag.
Nudansk Ordbog 2001 = *Politikens Nudansk Ordbog*. 18. udg. København: Politikens Forlag.
Nudansk Ordbog 2005 = *Politikens Nudansk Ordbog*. 19. udg. København: Politikens Forlagshus.
Ordbog over det danske Sprog 1919-1956. København: Gyldendalske Boghandel. Online: www.ordnet.dk
Talemåder i dansk 1998 = Stig Toftgaard Andersen: *Talemåder i dansk*. København: Munksgaard.

Christian Becker-Christensen
cand.mag.
Børglumvej 39
DK-2720 Vanløse
cbcvanlose@gmail.com

KONFERANSER

Rapport fra den 10. Konference om Leksikografi i Norden, Tammerfors, 3.-5. juni 2009

Marcin Overgaard Ptaszynski

Nordisk Forening for Leksikografi (NFL) holder en konference hvert andet år, og arrangementet omtalt her var den tiende i serien. Denne gang gik turen til Finland. Værtsbyen for konferencen var Tammerfors, Nordens største indlandsby, pragtfuldt beliggende mellem to søer (Näsijärvi og Pyhäjärvi). Den faglige del af konferencen foregik i auditorier på Tammerfors Universitet. Her samledes 104 deltagere, hovedsageligt fra de nordiske lande, men også fra Belgien, Tyskland og Litauen, for at præsentere deres ekspertise, viden og refleksioner om leksikografi.

Konferencen, arrangeret af Harry Lönnroth og Kristina Nikula, strakte sig over tre dage og bød på flere former for videnudveksling: plenarforelæsninger, foredrag arrangeret i to parallelle sessioner, posterpræsentationer og softwaredemonstrationer. Derudover, da konferencen blev delvist sponsoreret af forlaget WSOY og Forskningscentralen för de inhemska språken, var det muligt at se en udstilling af nyeste ordbøger med finsk.

De to første dage af konferencen begyndte hver med en plenarforelæsning. Den første af dem blev holdt af professor emeritus Reinhard Hartmann fra University of Exeter og handlede om hvorvidt leksikografi har opnået en status som en selvstændig akademisk disciplin. Adskillige argumenter, såsom udviklingen af leksikografisk teori, brugen af forskningsmetoder og egne diskursfora, kunne overbevise publikum om at leksikografien er blevet en akademisk disciplin. Den er *sui generis*, men samtidig rummer den både tværfaglighed og potentiale for videreudvikling. Den an-

den plenarforelæsning, givet af professor Henning Bergenholtz fra Aarhus Universitet, betragtede ligeledes leksikografien fra den teoretiske synsvinkel. Med udgangspunkt i den antagelse at en ordbog er en brugsgenstand, berettede foredragsholderen om nødvendigheden af at udarbejde ordbogen således at den tjener sit specifikke formål og dermed er tilpasset brugerens behov. Indholdet i begge foredrag er måske ikke noget nybrud i leksikografi; det er, i en eller anden form, blevet præsenteret af samme forelæsere ved tidligere lejligheder. På den anden side er begge emner helt centrale for leksikografien og dens udvikling, og alene af den grund var det en fordel at de blev taget op.

Der var ikke noget overordnet tema for konferencen, og det gav mulighed for at de 49 sessionsforedrag tilsammen kunne afspejle leksikografisk diversitet. Nogle emner var dog betydeligt mere populære end andre, og her kan man nævne følgende: flerordsforbindelser, dialektordbøger, arbejdet med kildemateriale og genbrug af ordbogsmodeller og databaser. Udvalgte bidrag der berørte disse emner, nævnes kort nedenfor.

Flerordsforbindelser af adskillige slags – idiomer, kollokationer, ordsprog eller sammensatte verber – var uden tvivl det mest populære emne ved konferencen. Det blev drøftet i ikke mindre end 11 foredrag, hvoraf 5 vedrørte svensk. Sven-Göran Malmgren og Emma Sköldberg diskuterede fordele og ulemper ved at placere idiomer direkte under den betydning som de hører til, frem for at samle dem efter alle betydningsangivelser. Den rigtige løsning afhænger, ifølge foredragsholderne, af ordbogens brugerprofil og funktion. Placeringen, men også udvalget og præsentationen, af kollokationer i Svenska Akademiens ordbok (SAOB) var emnet for Bodil Rosqvists foredrag. Hendes navnesøster og pendant fra Norske Akademis ordbogsredaktion, Bodil Aurstad, reflekterede om arbejdet med flerordsforbindelser i NAOB. Anna Hannesdóttir, Jón Hilmar Jónsson og Sofia Tingsell beskrev de fordele som elektronisk bilingval leksikografi byder på med hensyn til til-

gang til data om idiomer. Mens man i traditionelle papirordbøger normalt søger idiomer under et af de ord de består af, giver det elektroniske medie mulighed for at søge fremmedsproglige ækvivalenter af idiomer ud fra bl.a. deres pragmatiske funktioner.

Dialektordbøger blev omtalt af flere deltagere fra Finland og Danmark. Blandt de førstnævnte var Tarja Korhonen, som præsenterede detaljer vedrørende redigering af *Suomen murteiden sanakirja* (ordbog over samtlige finske dialekter) og Caroline Sandström, der berettede om arbejdet med *Ordbok över Finlands svenska folkmål*. I begge ordbøgers tilfælde er der tale om ambitiøse dokumentationsprojekter, som har været undervejs i flere år, og som stadig er langt fra at afslutte rejsen gennem hele alfabetet. Det samme gælder danske *Ømålsordbogen*, som blev omtalt af Asgerd Gudiksen. I sit bidrag fokuserede hun på arbejdet med indsamling af kildemateriale til ordbogen, især i processens tidlige fase, som begyndte i mellemkrigstiden. Belæg for leksikografisk sprogbeskrivelse blev også diskuteret – dog fra et mere moderne tidsperspektiv – af bl.a. Åse Westås og Knut Karlsen, som beskrev udfordringer i at bruge internetudgivelser som kildemateriale. Da indholdet af hjemmesider er dynamisk, er det svært at arkivere det løbende, især i store mængder. Kravet om dette kan dog ikke omgås hvis internet-baseret materiale skal bruges som kilde til en videnskabelig dokumentationsordbog som *Norsk Ordbok*.

Genbrug af materiale er – på godt og ondt – et af leksikografiens karakteristiske træk, og derfor undrer det ikke at det også blev et af de populære emner ved konferencen i Tammerfors. Begge deltagere fra Belgien, Godelieve Laureys og Maritta Moisio, fortalte hvordan den nederlandske orddatabase RBN og de kontrastive hollandsk-danske og hollandsk-finske databaser kan bruges til at lave en finsk-hollandsk og en finsk-dansk ordbog. Udfordringer der ligger i sådanne projekter, især problemer med at skabe ækvivalens, er både forudsigelige og velkendte i leksikografi, så gode råd til at tackle dem er altid værd at høre på. Leksikografisk genbrug af en mindre

kendt slags blev drøftet af Henrik Lorentzen og Sanni Nimb, som berettede om problemer forbundet med at omdanne definitioner fra *Den Danske Ordbog* til formaliserede wordnetdefinitioner.

Endelig var der flere sessionsforedrag som handlede mere om leksikologi end leksikografi. Olaf Almenningen, for eksempel, fortalte om det stigende antal anglicismer i norsk fodboldsprog, mens Lars Brink undersøgte pålydende betydning på flere sprog. I nogle tilfælde var der tale om en leksikologisk analyse foretaget på baggrund af et leksikografisk værk. Således handlede Torben Arboes foredrag om tryk i faste ordforbindelser ud fra data i *Jysk Ordbog*, mens Sven Lange undersøgte forandring af svensk ordforråd på baggrund af to ældre svenske ordbøger.

Sessionsforedragene blev så vidt som muligt grupperet efter emne og sprog. Ved mange konferencer skaber sådan en opdeling problemer for deltagere som gerne vil flytte midt i en foredragssession for at være med til en forelæsning i en parallel session. Takket være arrangørernes gode sans for organisation opstod problemet ikke ved denne konference. Sessionerne fandt sted i to auditorier der lå ved siden af hinanden, og der blev givet god tid til at skifte lokale for dem der ønskede det. Hvad arrangørerne dog ikke kunne sørge for, men heller ikke på nogen måde kunne bebrejdes, var at et par foredrag måtte aflyses, eller at deres indhold blev ændret i allersidste øjeblik. Det sker selv ved de bedste konferencer.

Foredragssessionerne blev på konferencens første dag afsluttet med to posterpræsentationer og på anden dag med to softwaredemonstrationer. De førstnævnte blev holdt af deltagere fra Finland og Tyskland og handlede henholdsvis om arbejdet med en finsk-tysk pædagogisk valensordbog (Jarmo Jantunen et al.) og om metoder for undersøgelse af interaktion mellem tekst og billeder i illustrerede ordbøger (Kati Lampinen). Til softwaredemonstrationerne blev der vist to generelle systemer til ordbogsredigering: det danske ILEX (Jens Erlandsen) og det norske system bag Norsk Ordbok (Øyvind Eide et al.).

Alt i alt bør må man sige at leksikografi ved konferencen i Tammerfors stod i sprogets tegn. Foredrag der handlede om andre opslagsværker end dem der bruges til at løse sproglige problemer, var der ikke mange af. Blandt undtagelserne kan nævnes bidragene ved Patrick Leroyer og Marcin Overgaard Ptaszynski, som handlede henholdsvis om leksikografisk kalibrering af informationsværktøjer og om behovstilpasset datatilgang i nordiske internetbaserede encyklopædier. Det undrer måske lidt at sprogordbøger var stærkt overrepræsenterede ved konferencen. Man har nemlig bevist at det er muligt at fylde en hel temasektion af *LexicoNordica* (bind 14) med bidrag som handler om andre opslagsværker end sprogordbøger. Tiden er helt bestemt moden til at erkende at leksikografi er meget mere end sprog. Forhåbentligt bliver det afspejlet i kommende konferencer om leksikografi i Norden.

Ud over muligheden for formel faglig diskussion bød konferencens arrangører også på et socialt program. På første dag blev deltagerne inviteret til tavastlandsk middag og aftenhygge i en historisk bomuldsfabrik, der nu huser det lokale kulturcenter. Andendagen bød på en officiel modtagelse på Tammerfors Rådhus og en guided tour i Tammerfors. Turen blev foretaget i busser, og det af gode grunde: Vejrguderne havde åbenbart glemt konferencen og sendte øsende regn og bidende kulde, med en junitemperatur der var rekordlav for de sidste 60 år.

Traditionen tro holdt NFL generalforsamling under konferencen. Blandt de vigtigste beslutninger var udskiftningen af redaktionskomiteen for *LexicoNordica* og valget af foreningens nye bestyrelse. Efter mangeårig tjeneste som hovedredaktører for *LexicoNordica* siden tidsskriftets første udgave bliver Sven-Göran Malmgren og Henning Bergenholtz afløst af Henrik Lorentzen og Ruth Vatvedt Fjeld. Den nye formand for NFL hedder Birgit Eaker. Hun afløser Halldóra Jónsdóttir på posten. Et udførligt referat fra generalforsamlingen kan hentes fra foreningens hjemmeside (http://www.nordisk-sprakrad.no/nfl.htm).

En omfattende rapport fra konferencen, under redaktion af Harry Lönnroth og Kristina Nikula, udkom i 2010. Den indeholder artikler skrevet på baggrund af konferencebidrag fra de fleste deltagere.

Da konferencen i Tammerfors var den tiende i serien, kan man sige at arrangementet fejrede sin første runde fødselsdag. Det giver anledning til en lille refleksion. I Oslo i 1991, ved åbningen af den første konference om leksikografi i Norden, afsluttede Inge Lønning sin velkomsttale med følgende ord: "Kjære nordiske leksikografer, – dere behøver ikke å frykte for arbeidsløshet! Lykke til med den første i en forhåpentligvis lang rekke av nordiske kongresser for et livsviktig fagområde" (Lønning 1992). Gode ønsker går somme tider i opfyldelse, og det har dette også gjort. Siden 1991 er NFL-konferencen nået Norden rundt næsten hele to gange (det er Sverige og Færøerne som foreløbigt kun har været vært for arrangementet én gang). Den har fastholdt evnen til at tiltrække interesse af både praktiserende leksikografer og forskere i nordisk leksikografi. Flere af deltagerne i den første konference er stadig aktive og har været med til konferencen i Tammerfors. De mangler ikke arbejdsopgaver, og hvert år får de nye, yngre kollegaer som deler deres iver i og fascination ved at udvikle ordbøger og andre informationsværktøjer for at lette kommunikation og videnstilegnelse. Det kan kun bekræfte at leksikografien fortsat er et livsvigtigt fagområde.

Det næste arrangement i serien, 11:e Konferensen om lexikografi i Norden, bliver afholdt 24.-27. maj 2011 i Lund i Sverige, med Svenska Akademiens ordbogsredaktion som vært. Man kan læse mere om konferencen på Nordisk Språkråds hjemmeside (http://www.nordisk-sprakrad.no/Konferanser.htm) og i dette nummer af *LexicoNordica*.

Litteratur

Lønning, Inge 1992: Ordets makt. I: Fjeld, Ruth Vatvedt (red.): *Nordiske studier i leksikografi. Rapport fra Konferense om leksikografi i Norden 28.-31. mai 1991*. Oslo: Nordisk Forening for Leksikografi.

Lönnroth, Harry/Kristina Nikula (red.) 2010: *Nordiska studier i lexikografi 10. Rapport från Konferensen om lexikografi i Norden, Tammerfors 3-5 juni 2009*. Tammerfors: Nordisk Forening for Leksikografi.

Marcin Overgaard Ptaszynski
adjunkt, ph.d.
Center for Leksikografi
Handelshøjskolen i Århus
Aarhus Universitet
Fuglesangs Allé 4
DK-8210 Århus V
maop@asb.dk

Inbjudan till 11:e Konferensen om lexikografi i Norden, Lund, 24–27 maj 2011

Svenska Akademiens ordboksredaktion, Nordiska föreningen för lexikografi och Språkrådet i Norge har nöjet att inbjuda till den 11:e Konferensen om lexikografi i Norden, som äger rum i maj 2011 i Lund, Sverige. Vi ser fram emot att träffa många nordiska och längre utifrån kommande lexikografer i en av Sveriges äldsta universitetsstäder.

Staden Lund präglas till stor del av universitetet som grundades 1666 och har drygt 40 000 studenter och fler forskare än något annat skandinaviskt universitet. Det stora antalet studenter bidrar till den ungdomliga stämningen och framtidspulsen i staden. Den centrala stadsbilden visar ännu idag på många historiska inslag med vindlande kullerstensgator, korsvirkeshus och stadsvallen från 1100-talet som omger delar av stadskärnan. De korta avstånden och den lugna trafiken har gjort Lund till en utpräglad cykelstad. Mitt i staden finns parken Lundagård, med bland annat domkyrkan, universitetshuset, Akademiska Föreningens borg och Palaestra et Odeum där konferensen kommer att äga rum.

Vetenskapligt program

Målet med det vetenskapliga programmet är att ge en så bred bild av den nordiska lexikografin som möjligt. Vi vill också ge deltagarna möjlighet att dryfta gemensamma frågor och få impulser utifrån. På konferensen hålls plenarföredrag av John Simpson, chef för Oxford English Dictionary, Godelieve Laureys, professor i skandinavistik i Gent, och Bo Ralph, professor i nordiska språk,

Göteborg. Vi erbjuder deltagarna möjlighet att hålla sektionsföredrag. Alla föredrag med anknytning till lexikografi är välkomna. Vi ser också gärna deltagare som presenterar posters och håller softwaredemonstrationer. Det kommer även att finnas möjlighet att presentera nyutkommen litteratur. För sektionsföredragen reserveras 20 minuter + 10 minuter för frågor och diskussion.

Under konferensen håller *Nordiska föreningen för lexikografi* också sin generalförsamling.

Socialt program

Konferensen öppnas på kvällen tisdagen 24 maj med en informell mottagning i Ordbokens hus i Lund. Man kan registrera sig på tisdagen eller på onsdagen. Konferensmiddagen hålls några mil utanför Lund på 1500-talsslottet Svaneholm torsdagen 26 maj och kombineras då med en bussutflykt i södra Skåne. Vi räknar också med en guidad rundvandring i centrala Lund.

Att resa till Lund

Det är lätt att ta sig till Lund från hela Norden. I närheten av Lund finns två flygplatser, Malmö Airport och Copenhagen Airport. Från Malmö Airport går det dagligen flera flygbussar direkt till Lunds city och bussfärden tar cirka 35 minuter. Från Copenhagen Airport går direkttåg till Lund och det tar cirka 45 minuter.

Tidsfrister

1.11.2010 Anmälan av föredrag, poster, workshop eller softwaredemonstration. En sammanfattning på högst 300 ord skickas i Word-format i en bifogad fil till Birgit Eaker, SAOB (birgit.eaker@svenskaakademien.se).

1.1.2011 Besked om antagna föredrag, posters, workshops och softwarepresentationer.

1.2.2011 Anmälan utan föredrag till reducerat pris.

Anmälningsformulär och mer detaljerad information om konferensen finns på NFL:s hemsida http://www.nordisk-sprakrad.no/nfl.htm.

Väl mött i Lund 2011!

Arrangörerna

REDAKSJONELT

1. LexicoNordica **udkommer** hvert år i november. Tidsskriftet indeholder leksikografiske bidrag som er skrevet på et af følgende nordiske sprog: dansk, finsk, færøsk, islandsk, norsk (bokmål eller nynorsk), svensk. Bidrag på engelsk, fransk eller tysk kan også optages hvis særlige forhold taler for det.

2. **Bidrag** sendes til det medlem af redaktionskomitéen som bor i bidragyderens land:

   Sturla Berg-Olsen, Norsk Ordbok 2014, Institutt for lingvistiske og nordiske studier, Universitetet i Oslo, Postboks 1021 Blindern, NO-0315 Oslo. sturla.berg-olsen@iln.uio.no

   Ken Farø, Københavns Universitet, Institut for Engelsk, Germansk og Romansk, Njalsgade 128, DK-2300 København S. kenfaroe@hum.ku.dk

   Jón Hilmar Jónsson, Stofnun Árna Magnússonar í íslenskum fræðum, Háskóla Íslands, Neshaga 16, ÍS-107 Reykjavík. jhj@lexis.hi.is

   Nina Martola, Forskningscentralen för de inhemska språken, Berggatan 24, FI-00100 Helsingfors. nina.martola@focis.fi

   Emma Sköldberg, Lexikaliska institutet, Institutionen för svenska språket, Box 200, SE-405 30 Göteborg. emma.skoldberg@svenska.gu.se

   Seneste tidspunkt for aflevering af bidrag er **den 1. april** hvis artiklen skal kunne trykkes i det nummer af tidsskriftet som udkommer i november samme år. Bidraget indleveres i elektronisk form, fx i Word-, Open Office- eller RTF-format.

3. **Illustrationer** der skal medtages i artiklen, indsættes i manuskriptet og vedlægges som separate grafikfiler.

4. **Manuskriptets ydre form**: Bidraget bedes forfattet i LexicoNordicas stilark, der kan rekvireres ved henvendelse til redaktionen. Manuskriptet **indledes** med forfatternavn og titel på artiklen. For tematiske og ikke-tematiske bidrag følger et **abstract** på engelsk på maks. ti linjer og dernæst selve artiklen, som opdeles i kapitler. Bidraget afsluttes med angivelse af post- og e-mail-adresse. Bidrag kan normalt have et omfang på højst 20 sider.

5. **Citater**: kortere citater (op til 3 linjer) bringes som en del af teksten med dobbelte anførselstegn omkring, mens længere citater eller fremhævelser af større vigtighed bør gives i et afsnit for sig selv uden anførselstegn.

6. Vi anbefaler en tilbageholdende brug af **fodnoter**. Evt. nødvendige noter gennemnummereres i teksten med højtstillet angivelse uden parentes.

7. **Litteraturhenvisninger** foretages i teksten på følgende måde:

   Herbst (2009:158) eller Herbst (2009).

   I den løbende tekst angives ikke hele internetadresser, men et forfatternavn eller en angivelse af titlen på internetbidraget, som bruges i litteraturlisten. Her angives internetadresser uden understregning.

8. Særlige angivelser: **leksikografiske termer** kan, når de indføres, fremhæves med fed; angivelse af **objektsproglige enheder** med kursiv, fx: ordet *ungkarl* har synonymet *alenemand*; betydninger af sproglige enheder angives ved hjælp af enkle anførselstegn, fx: 'en ugift mand'; dobbelte anførselstegn bruges ved

citater eller forbehold, fx: De er vokset op i de "glade" tressere. Tegnsætningsreglerne er forskellige i de nordiske lande, og artiklerne følger reglerne for det sprog som i øvrigt bruges i artiklen.

9. **Eksempel på litteraturangivelser**

Ordbøger

ALD 1948 = A.S. Hornby/E.V. Gatenby/H. Wakefield: *A Learner's Dictionary of Current English.* London: Oxford University Press.

COBUILD 1987 = *Collins* COBUILD *English Language Dictionary.* Editor in Chief: John Sinclair, Managing Editor: Patrick Hanks. London/Glasgow: Collins.

Hällström, Charlotta af & Mikael Reuter: *Finlandssvensk ordbok.* (1. upplagan). Helsingfors: Schildts 2001.

Norstedts stora engelska ordbok. Stockholm 2000.

Oxford-Hachette French Dictionary. Oxford: Oxford University Press 1994.

Anden litteratur

Haiman, John 1980: Dictionaries and Encyclopedias. I: *Lingua 50*, 329–357.

Lakoff, George og Mark Johnson 1980: *Metaphors we live by.* Chicago and London: The University of Chicago Press.

Mugdan, Joachim 1985a: Grammatik im Wörterbuch: Wortbildung. I: Herbert Ernst Wiegand (udg.): *Studien zur neuhochdeutschen Lexikographie IV*. Hildesheim/Zürich/New York: Olms, 237–308.

Zgusta, Ladislav 1971: *Manual of lexicography*. The Hague: Mouton.

Internethenvisninger

ELEXIKO = Klosa, Annette m.fl. (red.): *elexiko.* Mannheim: Institut für Deutsche Sprache. www.elexiko.de (maj 2008)

STO = Braasch, Anna m.fl. (red.): *Sprogteknologisk Ordbase.* København: Center for Sprogteknologi 2001-2004. www.cst.dk/cgibin/defisto (april 2007)

Finin, Tim 2006: On evaluating the credibility of Wikipedia articles. http://ebiquity.umbc.edu/blogger/2006/11/23/on-evaluating-wikidedias-credibility/ (februar 2007)

10. LexicoNordica vil udkomme både som trykt tidsskrift og i en **internetudgave**. Ved indsendelse af et bidrag til redaktionen erklærer forfatterne sig derfor indforstået med en elektronisk udgivelse.