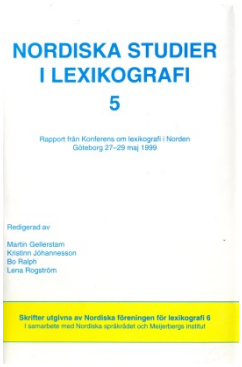
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Lärdomar från utveckling av inflekterande synonymordböcker

Forfatter: Antti Arppe

Kilde: Nordiska Studier i Lexikografi 5, 2001, s. 31-44
Rapport från Konferens om lexikografi i Norden, Göteborg 27.-29. maj 1999

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



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

Antti Arppe

Lärdomar från utveckling av inflekterande synonymordböcker

During 1996–1998, product development projects were carried out at Lingsoft in order to create so-called inflecting thesauri, in which the electronic form of the contents of synonym dictionaries for Swedish, Danish, and Norwegian Bokmål were integrated with computerized morphological models, created according to the two-level model, for the respective languages. The resultant computer programs provided practical insights into the interaction between structures of semantic relations, in this case representing synonymity, and with inflectional morphology. As a result, it became evident that the principle of lexical generality cannot be trusted blindly in generating across the board the inflected forms of the base-form components of a synonym dictionary. Furthermore, it would seem that synonymity between words cannot be categorically expected to extend throughout the entire inflectional paradigm of these words. This would suggest that inflected forms of words should also be considered when constructing structures of semantic relations.

1. Introduktion

Denna artikel bygger på erfarenheter från produktutvecklingsarbete av flera s.k. inflekterande synonymordböcker, som utfördes på Lingsoft 1996–1998. Dessa program utvecklades för de främsta nordiska skriftspråken i ordningen svenska (färdig 1996), danska samt norskt bokmål (1997). Först kunde dessa program bara brukas i Windows-operativsystem, men Macintosh-versioner av alla blev färdiga efteråt (1998). Redan tidigare hade man producerat ett likadant verktyg för finska (1995–1996), och nästan samtidigt med de skandinaviska språken fortsatte vi samma arbete för tyska (1997–1999), men dessa berörs inte i denna artikel.

Här skall jag först beskriva vad en inflekterande synonymordbok är och hur vi utvecklade dessa program. Därefter skall jag berätta om de lingvistiska insikter som vår produktutvecklingsgrupp fick under arbetets gång. Denna artikel bygger på en

tidigare arbetsrapport som jag författat tillsammans med Mari Voipio och Malene Würtz (Arppe, Voipio, Würtz 1999). Därtill är jag tacksam för synpunkter och kommentarer från alla andra som deltog i dessa produktutvecklingsprojekt, särskilt Jussi Birn som deltog i det svenska delprojektet.

2. Inflekterande synonymordböcker och deras utvecklingsarbete

Elektroniska ordböcker av varierande kvalitet och ambition har redan funnits hela 1990-talet för de främsta nordiska skriftspråken. Hittills har nästan alla innehållit ord bara i grundform, och har alltså mer eller mindre varit elektroniska kopior av verk i bokform, med den fördel som man får vid sökning i elektroniska datastrukturer. Om man börjar i en sådan elektronisk synonymordbok med någon böjd form av ett ord, måste man först analysera ordet självt för att mata in dess grundform i programmet. Om man därefter vill få någon synonym i samma böjningsform som det ursprungliga ordet har, måste man göra också detta själv via tangentbordet. En inflekterande synonymordbok har till uppgift att automatisera denna process så att ordboksprogrammet kan behandla inte bara ord i grundform utan också alla deras böjningsformer. Denna egenskap ökar synonymordboksprogrammets användbarhet när man skriver eller redigerar löpande text – nyttigt i alla språk som har något böjningssystem alls, vilket gäller alla nordiska språk och särskilt finska.

Schematiskt sett fungerar en inflekterande synonymordbok på följande sätt:

I morfologiskt analys av ett böjt ord eller en fras
(⇐ användaren)
⇒ ordets eller frasen grundform eller -former
⇒ ordets eller frasens böjningsinformation

⇓

II sökning av synonymer från semantiska datastrukturen med det ursprungliga ordets grundform (⇐ steg I)
⇒ lista av synonymer i grundform

⇓

III generering av respektive böjningsformer av synonymorden
(⇐ steg II) eller -fraser enligt böjningsinformation (⇐ steg I)
⇒ lista av synonymer i respektive böjningsform(er)

Som ett konkret exempel på hur detta fungerar i praktiken kan man titta på följande synonymer av det svenska ordet *klassificerades*. Det kan analyseras antingen som ett verb i passiv preteritumform, som t.ex. i *boken* ***klassificerades***, eller som participet i genitivform, som t.ex. i *de* ***klassificerades*** *ordning* – båda av verbet *klassificera*. I många fall är synonymerna organiserade enligt källan i två eller flera grupper, vilket framgår nedan av verbsynonymerna:

klassificerades
⇒ ordnades (verb)
delades upp i klasser, indelades, klassades, grupperades, rangordnades, rankades, inrangerades, graderades, kategoriserades, systematiserades
⇒ betecknades (verb)
bedömdes
⇒ betecknades (particip)
bedömdas

Som råvara för en sådan inflekterande synonymordbok behöver man två grundkomponenter. Den första är en datalingvistisk morfologisk modell av språket – som i vårt fall var tvånivåmodellen enligt Koskenniemi (Koskenniemi 1983). Den danska tvånivåmodellen hade licensierats tidigare från Thomas Bilgram (Bilgram 1994) och vidareutvecklats på Lingsoft, men modellerna för svenska och norskt bokmål hade utvecklats enbart på Lingsoft (Karlsson 1992, Moshagen 1999). Den andra komponenten är en elektronisk version av en synonymordbok, som vi har licensierat från välkända nordiska förlag utom för finska, som hade utvecklats internt på företaget tidigare. Det svenska synonymordboksinnehållet, med Göran Walter som författare, licensierades från Walters lexikon, vilket material hade publicerats i bokform som *Bonniers synonymordbok* av Bonnier Alba 1995. Det danska synonyminnehållet, med Allan Karker som författare, licensierades från Politikens Forlag, och hade publicerats som *Politikens Synonymordbog* år 1994. Det norska synonyminne-

hållet, med Herbert Svenkerud som författare, licensierades från J.W. Cappelens Förlag, och hade publicerats som *Cappelens Media Synonymordbok* redan år 1983. Alla dessa synonymordböcker skilde sig från varandra i uppläggning, men passade med mindre eller större omarbetning till slutprodukten. Det svenska materialet påminde kanske mest om en traditionell synonymordbok med några tiotusen uppslagsord och en eller flera grupper av synonymer för varje uppslagsord, medan det danska materialet nästan helt saknade gruppering av synonymer, och det norska materialet var grupperat under flera tusen grunduppslagsord, som innehöll långa listor av närmare och avlägsnare synonymer – lyckligtvis semantiskt grupperade.

Utvecklingen av varje inflekterande synonymordbok tog flera månader, av vilka den största delen utgjordes av omarbetning av det ursprungliga materialet, från t.ex. enkla ordbehandlingsfiler, till en systematiskt strukturerad elektronisk form med alla förkortningar utskrivna, och därefter i testandet och finjusteringen av det resulterande programmet. I varje projekt deltog en programmerare, en eller flera lingvister samt en projektchef. Den svenska versionen utvecklades helt under året 1996 och kom ut på marknaden som en färdigintegrerad modul i den svenska versionen av Microsoft Office 97. Både den danska och norska bokmålsversionen utvecklades under 1998 och blev en del av uppgraderingsversionen 97A av respektive danska och norska Microsoft Office 97.

3. Erfarenheter med lingvistisk betydelse

3.1. Lexikal allmängiltighet

Teorin i generativ morfologi påstår att man kan anta att varje ord som hör till en viss ordklass kan böjas i alla former i ordklassens böjningsparadigm. Bl.a. Scalise (1984) säger följande:

> It is in general possible to attach to any word the entire set of inflections associated with the word [base form] in question.

Bybee (1985) kallar detta för *lexikal allmängiltighet* (← *lexical generality*). Detta gäller särskilt när man tänker på nya ord, som man kan anta bli böjda på samma sätt som de andra, existerande medlemmarna i samma ordklass som de nya orden, även om dessa inte existerar (dvs. inte har använts) än. T.ex. kan man anta att alla svenska utrumsubstantiv som slutar på *-a* böjs i alla böjningsformer enligt det traditionella mönstret, vad än dessa ord betyder: *en rucka, ruckan, ruckor, ruckorna, ruckas, ruckans, ruckors, ruckornas*. Detta är en egenskap hos de öppna ordklasserna. Däremot gäller lexikal allmängiltighet inte vid derivation. Om man kan derivera ett substantiv som *tävling* från verbet *tävla*, garanterar detta inte att man kan göra likadant med verbet *finna*. I stället för **finning* bör man använda en annan derivationsändelse, t.ex. *finnande*. Från denna synpunkt kan man se datoriserade morfologiska modeller av något språk som en egentlig inkarnation av lexikalisk allmängiltighet, då dessa byggs upp av ett lexikon med ordstammar som klassificeras enligt ordklass, och en regelsamling enligt vilket ord tillhörande en viss ordklass böjs.

På det sättet kan man beskriva en inflekterande synonymordbok som en tillämpning av lexikalisk allmängiltighet över hela innehållet av en synonymordbok. Vad som då händer i stor skala har inte kunnat observeras tidigare, då det ännu inte funnits både datoriserade morfologiska modeller och synonymordböcker i elektronisk form. Vi förväntade oss att en sådan generering av former skulle behöva, lingvistiskt sett, bara minimal korrigering, men realiteten var annorlunda. Det uppstod ett behov av både *morfosemantiska* och *pragmatiska* begränsningar i det ursprungliga lexikonet.

Morfosemantiska begränsningar kallas här de böjningsformer som en infödd språkbrukare helt enkelt tolkar som felaktiga. T.ex. kan en del verb egentligen inte användas i passivform, som svenska **vars* och **blevs* eller norska **gikks*. Därtill kan en del adjektiv inte få både komparativ- eller superlativform, t.ex. danska *kæmpestor, skidegod* och *ekstra,* men inte **kæmpestørre*, **skidebedre* och **ekstrare* eller **kæmpestørst*, **skidebedst* och **ekstrast*. Detta är först och främst en semantisk, ordspecifik

bedömning, och därför var vi tvungna att genomgå de ursprungliga lexikonens alla verb och adjektiv för att klassificera varje ordstam i dessa ordklasser i detta avseende (i substantiv kunde man välja ut ett begränsat antal pluralia tantum-fall). Orsaken till detta var antagligen främst att i morfologisk analys, till vilket ändamål våra morfologiska modeller ursprungligen var utvecklade, behöver beskrivningen inte vara så normativt noggrann som vid morfologisk generering, vilken egenskap vi behövde i våra inflekterande synonymordböcker. Så de morfosemantiska begränsningarna var egentligen luckor i de ursprungliga morfologiska modellerna. Men i varje fall gäller sådana noggrannare klassifikationer och böjningsbegränsningar slutligen bara de specifika ord som har klassificerats, inte alla deras möjliga derivationer eller sammansättningar med andra ord.

Behovet av morfosemantiska begränsningar är också en begreppsdefinitionsfråga. Om man räknar som ordklasser bara de knappt tio som man traditionellt lär sig i skolan, substantiv, adjektiv, verb osv., blir man tvungen att efteråt granska genereringen av olika böjningsformer av ord tillhörande varje böjd ordklass, eftersom alla ord i dessa ordklasser inte uppvisar alla böjningsformer. Däremot, om man definierar en ordklass som en grupp ord som har samma böjningsparadigm, med likadana begränsningar eller luckor i paradigmet, och bygger lexikonet enligt denna strängare princip, finns den produktiva begränsningen färdig i lexikonet, och behovet av ytterligare granskning försvinner. När man tar hänsyn till alla kända undantag i mönsterparadigmet, har vi när det gäller svenska kommit fram till mer än 200 olika variationer – man borde säga miniordklasser.

Pragmatiska begränsningar gjordes för några böjningsformer som bedömdes av en infödd språkbrukare som konstiga, även om de inte var helt ogrammatiska. Som sådana klassificerades passiva preteritumformer av vissa verb, som danska *?bandtes* och *?lodes* (av *binde* och *lade*). Likaså skulle imperativformen av några verb låta konstiga, t.ex. danska *eksistere – ?eksister!* eller norska *sykle – ?sykl!* Som pragmatiska begränsningar räknades också de fall då en viss analys, och det därtill korresponderande genererade ordet, skulle vara väldigt sällsynt och därför inte för-

väntat. Sådana var genitivform av adjektiv i singularis, t.ex. svenska *?stors, ?storts* och de motsvarande participformerna, t.ex. danska *?spists, ?elskets*. Likadana bedömningar gjordes för komparativ- och superlativformer av vissa adjektiv. Medan några adjektiv som betecknar färg lätt kan kompareras, t.ex. svenska *grön – grönare – grönast*, skulle andra låta konstiga, t.ex. danska *orange – ?orangere – ?orangest*.

Pragmatiska begränsningar behövdes också på grund av homonymi, som förekommer relativt ofta i de skandinaviska språken. T.ex. danska *flyver* kan analyseras både som ett verb i presensform och som ett substantiv i grundform, och svenska *anklagades* antingen som passiv preteritum av verbet *anklaga* eller som den bestämda definita genitivformen av participet *anklagad*. Om man hade ordens kontextinformation, så kunde man lätt skilja mellan dessa alternativa analyser med existerande programverktyg. Problem uppstår när de existerande gränssnitten med ordbehandlingsprogram inte erbjuder någon kontextinformation. Detta är inte särskilt problematiskt då de båda analyserna är relativt frekventa och brukaren själv kan välja det alternativ som gäller. Men när någon analys är relativt infrekvent valde vi att begränsa dessa former för att minska brukarens förvirring, särskilt eftersom många inte betraktar dessa ord genom en lingvists glasögon. Man kan anta t.ex. att svenska *gås* hänvisar till fågeln och inte är presens passiv form av verbet *gå*.

3.2. Synonymi och böjning

Den föregående delen av denna uppsats har behandlat skillnader i böjningen av enstaka ord inom olika ordklasser, som först och främst är en fråga begränsad till den morfologiska analys- och genereringskomponenten av en inflekterande synonymordbok. När man kombinerar en semantisk relationsbeskrivning, i detta fall innehållet i en synonymordbok, med en sådan automatisk böjningsapparat, mötte vi ett svårare problem: Hur "bra" kommer det synonyma förhållandet mellan två eller flera ord att kvarstå i de respektive ordens olika böjningsformer?

Den nuvarande lexikografiska och semantiska teorin behandlar inte denna fråga – åtminstone inte direkt. Allmänt verkar böjningens interaktion med och påverkan på ordens semantiska egenskaper ha tillmätts en underordnad betydelse. T.ex Zgusta tonar ned böjningen som bara ett av många olika drag som ett ords betydelse består av, och Cruse förklarar böjningen som ett sekundärt drag semantiskt sett, som inte ändrar grundformens centrala betydelse:

> The lexical meaning comprises of a great number of different components and there are many pertinent phenomena and relations which must be studied completely, but preferably as distinct, interdependent factors, if the whole meaning of a word (lexical unit) is to be analyzed, understood, and described. It will be useful to discern the following main components of lexical meaning: 1) the **designation**, 2) the **connotation**, and (possibly) 3) the **range of application**. (Zgusta 1971: 27)

> We have so far assumed that it is a word form associated with a single sense, and that a difference of word form entails a difference of lexical unit. [...] Strictly speaking, we would be obliged, on this view, to regard, for instance *obey*, *obeys* and *obeyed* as representing different lexical units. It would, however, be more advantageous for our purposes to be able to say that they were **alternative manifestations** of the **same lexical unit** *obey*. (Cruse 1986: 76–77)

Senare, efter diskussion av grundkomponenterna i ordets betydelse, nämner Zgusta dock böjningens betydelse, och påminner om den variation i betydelse inom ett ords böjningsparadigm som påträffats i flera språk:

> It cannot be stressed enough that in different languages, any members of any paradigm, whether 'regular' or 'irregular', can have a lexical meaning more or less different from the rest of the paradigm. [...] In any case, the lexicographer should always count on the possibility that a form – any form – of a paradigm can show a peculiarity in respect to its lexical meaning when compared with the other forms of the paradigm. (Zgusta 1971: 123, 126)

Trots denna observation, som Zgusta stöder med flera exempel, har han inte valt att betrakta böjningen som en av ordbetydelsens

huvudkomponenter. Senare har Sinclair också antytt detsamma utgående från sina erfarenheter i engelsk lexikografi:

> ... it is not yet understood how meanings are distributed among forms of a lemma, and a new branch of study is looming – the interrelationships of a lemma and its forms. (Sinclair 1991: 41)

När man fortsätter i litteraturen till beskrivning av ordens semantiska förhållanden verkar böjningens semantiska betydelse försvinna helt och hållet. I fall av synonymi koncentrerar man sig på ordens likhetsgrad, men antyder ingenting om böjning. Inte heller Zgusta tar upp böjningen när han beskriver synonymi:

> two words are synonymous, if they have the same sense. (Lyons 1968)

> Synonyms, then, are lexical items whose senses are identical in respect of 'central' semantic traits, but differ, if at all, only in what we may call 'minor' or 'peripheral' traits. (Cruse 1986: 267)

> It is, however, advantageous to reserve the term synonymy, synonyms [...] only for the cases of absolute identity in meaning. [...] The absolute identity of meaning requires the identity of all the three basic components of meaning [being designatum, connotation and range of application]. [...] If there is a difference at least in one of the three basic components of meaning [...], the respective words are near-synonyms only. (Zgusta 1971: 89, 90)

Teorin har också följts i praktiken. I WordNet har man beslutat att lämna den morfologiska böjningen utanför den centrala semantiska beskrivningen:

> Initially, interest was limited to semantic relations; no plans were made to include morphological relations in WordNet. [...] English nouns, verbs, and adjectives are organized in synonym sets, each representing one underlying lexical concept. [...] The three papers [describing the principles of WordNet] have little to say about lexical relations resulting from inflectional morphology, since those relations are incorporated in the **interface** to WordNet, **not in the central database**. (Miller 1993: 9)

I vårt testande av de första versionerna av inflekterande synonymordböcker märkte vi att det inte i praktiken går att bortse från böjningens påverkan på ordens synonymiförhållanden. Särskilt

var verbsynonymer känsliga för böjningen. Låt oss titta på de svenska synonymparen *hitta–finna* och *säga–betyda*. Fast dessa ord är synonyma i meningarna (1–4), är det inte längre möjligt att använda dem som synonymer för varandra i meningarna (5–8):

(1) Sen *hittade* jag boken.
(2) Sen *fann* jag boken.

(3) Vad vill det *säga*?
(4) Vad vill det *betyda*?

(5) Boken *hittades*.
(6) Boken *?fanns*.

(7) *Säg* någonting!
(8) **Betyd* någonting!

I dessa exempel försvinner det synonymiska forhållandet antingen på grund av att ordet ändrar sin semantiska betydelse i en eller flera böjningsformer (6), eller att den korresponderande böjningsformen är ogrammatisk för synonymordet (8). För svenskans del fann vi att denna inkongruens återfanns relativt ofta i verbsynonymgrupper och särskilt uppstod vid passiv- och imperativform, som vi därefter var tvungna att granska manuellt i detta avseende. Likadana fenomen fanns också i de danska och norska inflekterande synonymordböckerna.

Sammantaget var vår slutsats att man inte kategoriskt kan förvänta att synonymiförhållanden mellan två eller flera ord (i grundform) automatiskt också kvarstår i alla böjningsformer av dessa ord. En orsak till att detta fenomen inte behandlats särskilt mycket i litteraturen är kanske att hela idén att generera synonymbeskrivningar automatiskt inte riktigt har uppstått tidigare. Om man utgår från enbart de tryckta synonymordböckerna, är frågan inte så problematisk, för de synonymrelationer som beskrivs i ordboken gäller ju bara de former som finns tryckta i ordboken. I praktiken garanterar man ingenting beträffande böjda former av dessa ord, men psykologiskt sett pekar valet av nästan bara grundformer i sådana verk starkt i den riktningen att dessa semantiska förhållanden också skulle kvarstå i grund-

formernas böjda former. Men om man trovärdigt vill ange synonymer till verbens passivformer borde man egentligen skapa en separat synonymordbok för detta ändamål, utgående från varje böjningsform.

4. Sammanfattning

Vårt arbete för att åstadkomma inflekterande synonymordböcker för danska, norskt bokmål och svenska genom att integrera datoriserade morfologiska modeller för respektive språk med elektroniska versioner av synonymordböcker har gett betydande insikt i flera lingvistiska fenomen, särskilt i skärningsplanet mellan lexikalisk morfologi och lexikalisk semantik, mer specifikt synonymi. De viktigaste lärdomarna är följande:

1. Lexikalisk allmängiltighet kan man inte lita på i den utsträckning som man skulle anta, särskilt om man bara använder en grov ordklassindelning. Man behöver både morfosemantiska, morfologibaserade begränsningar i modellens ordformsgenerering och pragmatiska begränsningar på grund av frekvens- eller förväntningsinformation.
2. Man kan inte kategoriskt förvänta att synonymiförhållanden mellan två eller flera ord (i grundform) automatiskt också kvarstår i alla böjningsformer av dessa ord.

Till slut verkar våra erfarenheter antyda att det i det här området finns många frågor att forska vidare i.

Kontaktinformation

Antti Arppe, Lingsoft Ab/Helsingfors Universitet
antti.arppe@iki.fi

Litteratur

A. ORDBÖCKER

Politikens Synonymordbog. 1994. Karker, Allan. Politikens Forlag.

Cappelens Media Synonymordbok. 1983. Svenkerud, Herbert. J.W. Cappelens Forlag.

Bonniers Synonymordbok. 1995. Walter, Göran. Bonnier Alba.

B. ANNAN LITTERATUR

Arppe, Antti; Voipio, Mari; Würtz, Malene. 1998. "Creating inflecting electronic thesauri." Working paper presented at the 17th Scandinavian Conference in Linguistics, August 20–22, 1998.

Bilgram, Thomas. 1994. "Computerstyret analyse af dansk. En praktisk analyse af en væsentlig kilde til homonymi i dansk med forslag til kontekstuell disambiguering ved hjælp af constraint-regler." Upubliceret speciale, Institut for Lingvistik, Aarhus Universitet.

Bybee, Joan L. 1985. *Morphology*. John Benjamins Publishing Company.

Cruse, D.A. 1986. *Lexical Semantics*. Cambridge: CUP.

Karlsson, Fred. 1992. SWETWOL: A Comprehensive Morphological Analyzer for Swedish. *Nordic Journal of Linguistics* 15:1–45.

Koskenniemi, Kimmo. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Helsinki: Publications of the Department of General Linguistics 11, Department of General Linguistics, University of Helsinki.

Lyons, John. 1968. *Introduction to theoretical linguistic*. Cambridge: CUP.

Miller, George A. et al. 1993. Introduction to WordNet: An On-Line Lexical Database. URL.
http://www.cogsci.princeton.edu/~wn/... (visited June 1999)

Moshagen, Sjur. 1998. "Morphological Analyzers for Norwegian (Bokmål and Nynorsk)." Working paper presented at the 17th Scandinavian Conference in Linguistics, August 20–22, 1998.

Scalise, Sergio. 1984. *Generative Morphology*. The Netherlands: Foris Publications.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Zgusta, Ladislav. 1971. *Manual of Lexicography*. (*Janua Linguarum, Series Maior*. 39.) The Hague: Mouton.