Att hålla på formerna
Om framväxten av Svensk morfologisk databas

Sture Berg & Yvonne Cederholm

Bakgrund

Sedan ett par år pågår vid Språkdata ett arbete i syfte att skapa en s.k. fullformsdatabas med utgångspunkt i den senaste upplagan av SAOL (1998). Den officiella benämningen på databasen i fråga är Svensk morfologisk databas (SMDB). Vi vill i denna artikel kort beskriva tillkomstsättet av och innehållet i databasen och samtidigt förmedla några tankar om vad den kan användas till i framtida lexikonarbete. Det bör tilläggas att SMDB är en vidareutveckling av projektet SAOL i maskinläsbar form, som bedrevs vid institutionen under senare delen av 1980-talet och finns beskrivet i Berg & Samuelson 1988.

Övergripande utgångspunkter och mål

I SMDB utgör uppslagsorden i SAOL12 utgångspunkten. Databasen lagrar emellertid inte bara uppslagsordens grundformer utan också i förekommande fall samtliga böjningsformer. För varje enhet ges dessutom ordledsmarkeringar enligt den modell för morfologisk ord-uppdelning som redovisas i bl.a. inledningen till SAOL och som tillämpas på materialet där. Alla ingående grafiska ordformer i databasen beskrivs även morfosyntaktiskt med hjälp av s.k. taggar (*hästar* får en märkning för obestämd form plural grundform till substantivet *häst*, medan *händer* får två olika märkningar, dels för den obestämda pluralformen till substantivet *hand*, dels för presensformen till verbet *hända*).

Kombinationen av grafisk form, tagg och lemmatillhörighet kallar vi för lemmatisk fullform.

Själva etableringen av böjningsformerna (som gått under arbetsnamnet uppblåsningen) bygger på det mer eller mindre självklara faktum att de ord i svenskan – och det är de flesta – som vid sidan av sina grundformer också uppvisar ett större eller mindre antal böjda former grupperar sig enligt vissa mönster, om än med många olika typer av avvikelser. Första fasen i projektet utgjordes sålunda av en komplett märkning av SAOL:s inemot 120 000 uppslagsord. Ordet *fisk* fördes till grupp 12 liksom alla sammansättningar med *-fisk* som efterled (*bläckfisk*, *guldfisk* etc.). I samma grupp hamnade också substantiv som t.ex *bunt*, *frack* och *yngling*. Gruppen konstitueras sålunda av att bestämd form singular slutar på *-en* och obestämd form plural på *-ar*. De önskade böjningsformerna kan alltså erhållas via enkla regler som kopplar ändelserna direkt till stammen. Ordet *dröm* däremot kunde inte tillföras samma grupp. Visserligen har vi att göra med samma ändelsemorfem som i fallet *fisk*, men eftersom stamkonsonanten måste dubbleras i samtliga böjningsformer (utom obestämd form singular) kom ordet i stället att tillhöra grupp 15.

Vid indelningen av det kompletta SAOL-materialet i enhetliga grupper måste vi alltså ta hänsyn till många viktiga variabler som flyktig vokal, konsonantfördubbling, omljud, avljud, frånvaro av pluralformer hos vissa substantiv etc. Sammantaget uppgår antalet olika grupper för närvarande till mer än 300. Inom parentes bör nämnas, att de olika gruppbeteckningarna är arbiträrt valda, och att omstruktureringar i det hierarkiska systemet kommer att företas i ett senare skede. Huvudsaken just nu är innehållet i grupperna och kopplingen av dem till de regler som genererar de lemmatiska formerna. Det bör också tilläggas, att utgångspunkten för indelningen i ’olika ord’ går tillbaka på den s.k. lemma-lexem-modellen (beskriven bl.a. i Allén 1970, 1971 och Berg 1978). Denna modell – som ju för första gången i SAOL-

sammanhang tillämpas i den senaste upplagan – har varit en förutsättning för uppbyggandet av SMDB enligt ovan.

Generering av lemmatiska fullformer

Genereringen utgår från böjningsgruppens nummer och från uppslagsordets grafiska form. För ordet *fisk* med gruppnummer 12 genereras substantivets samtliga åtta böjningsformer. Reglerna (en för varje böjningsform) anger vilket böjningssuffix som ska läggas till stammen (uppslagsordet). I förekommande fall anges i reglerna också stamoperationer som förändrar stammen på något sätt. Ordet *dröm* till exempel tillhör grupp 15. Regeln som genererar böjningsformen *drömmen* innehåller en stamoperation ”%dk” som dubblerar den sista konsonanten i stammen. Varje regel genererar också en morfosyntaktisk beskrivning (tagg) samt en variant av böjningsformen som innehåller ordledsmarkeringar (stora och små). Generingssystemet omfattar ca 4 000 olika regler. Exemplet nedan visar reglerna för böjningsgrupp 15.

TABELL 1. *Regler för böjningsgrupp 15.*

operation	tagg	*exempel*
=	NCUSNI	dröm
=+s	NCUSGI	dröms
%dk+/en	NCUSND	drömm+en
%dk+/ens	NCUSGD	drömm+ens
%dk+/ar	NCUPNI	drömm+ar
%dk+/ars	NCUPGI	drömm+ars
%dk+/ar/na	NCUPND	drömm+ar+na
%dk+/ar/nas	NCUPGD	drömm+ar+nas

De sexton stamoperationerna har definierats så att de tillför databasen information som kan vara intressant att söka på i ett senare skede. Exempelvis finns operationer för att hantera omljud, avljud och flyktig vokal. Vi har alltså inte i första hand strävat efter att hålla nere antalet regler, utan istället prioriterat möjligheten att söka i databasen efter t.ex. substantivformer

med flyktig vokal eller verbformer med avljud. En vanlig operation för verben är att radera sista vokal i stammen eftersom reglerna inte opererar på tekniska stammar som t.ex. *lek* för verbet *leka* utan på hela uppslagsordet. Orsaken är att inga tekniska stammar är avgränsade i SAOL. I det sista skedet tillämpas också vissa filter på de grafiska formerna. Ett sådant filter tar bort s-suffix på ord som slutar på ”s”, ”z” eller ”x”. Ett annat filter eliminerar vissa trekonsonantkombinationer i böjningsformerna. Om tre likadana konsonanter följer på varandra behålls samtliga konsonanter endast i formen med ordledsmarkeringar, alltså *tillåta* men *till=låta* (tecknet ”=” betecknar stor ordledsgräns).

Tagguppsättningen

Den morfosyntaktiska beskrivningen i SMDB ska så långt som möjligt överensstämma med böjningsinformationen i SAOL12. Vi har också valt att lägga oss nära beskrivningen i Nusvensk frekvensordbok (Allén 1970, 1971) när det gäller intern homografi inom ett lemma. Här urskiljs t.ex. tre olika böjningsformer på ”-a” hos adjektivet (utrum och neutrum i singular samt plural). Som en jämförelse kan nämnas att Stockholm Umeå Corpus Project/SUC (Ejerhed et al. 1992) arbetar med två former (singular och plural).

SMDB:s tagguppsättning grundas på EAGLES (1996) allmänna rekommendationer för morfosyntaktisk annotering. Vi har eftersträvat en tagguppsättning som i mycket hög grad liknar de tagguppsättningar som används inom andra projekt på institutionen (PAROLE, Kokkinakis & Kokkinakis Johansson 1997). Taggarna består av ett antal attribut-värdepar där positionen anger vilken kategori som avses. Den första positionen i taggen anger alltid ordklass. Övriga positioner/kategorier beror på vilken ordklass som beskrivs. För substantiv anger 2:a position typ, 3:e position genus, 4:e position numerus, 5:e position kasus och slutligen 6:e position bestämdhet (species). I exemplet nedan visas taggarna för

substantivets singulara former för ordet *dröm (*jämför för övrigt tabellen i föregående avsnitt).

NCUSNI	substantiv utrum singular nominativ obestämd (*dröm*)
NCUSGI	substantiv utrum singular genitiv obestämd (*dröms*)
NCUSND	substantiv utrum singular nominativ bestämd (*drömmen*)
NCUSGD	substantiv utrum singular genitiv bestämd (*drömmens*)

Långsiktigt underhåll av SMDB

En viktig poäng med SMDB-projektet är att se till att den morfologiska databasen uppdateras kontinuerligt. SMDB kommer att ha en nära koppling till SAOL och andra nuspråkliga lexikaliska databaser vid Språkdata. När en ny upplaga av SAOL är under arbete förses nytillkomna uppslagsord med böjningsgrupp, varigenom böjningsformerna automatiskt tillförs den morfologiska databasen. När lemman tas bort i SAOL försvinner de också i SMDB. Frekvensuppgifter uppdateras då nya korpusar integreras i Språkbanken. På så sätt får man ett system med en levande kontakt mellan texter och lexikon.

Annotering av Språkbanken

SMDB kommer att utgöra grunden i ett annoteringssystem för lemmatisering och morfosyntaktisk taggning av Språkbankens nuspråkliga korpusar. De annoterade korpusarna kommer att göras tillgängliga genom Språkbankens nya korpussystem (Gellerstam, Cederholm & Rasmark 2000).

Vid annoteringen förses varje graford med två slags taggar: den ena taggen är den morfogrammatiska beskrivningen och den andra anger lemmatillhörighet. Exempelvis annoteras grafordet *bilder* med den morfosyntaktiska taggen NCUPNI (substantiv, utrum, plural, nominativ, obestämd form) och lemmat *bild*.

Homografa former annoteras med flera analyser. Som illustrationsexempel tar vi fallen *barn, ropa, klöv* och *tår*. För *barn* ges analyserna 1) NCNSNI (substantiv, neutrum, singular, nominativ, obestämd form) samt 2) NCNPNI (substantiv, neutrum, plural, nominativ, obestämd form) och för *ropa* 1) V0N0A (verb, indikativ, infinitiv, aktiv) samt 2) V0M0A (verb, indikativ, imperativ, aktiv). Här exemplifieras **intern homografi** (två lemmatiska former med samma grafiska former, samma lemmabeteckningar men olika taggar). I fallet *klöv* får vi analyserna 1) V0ISA (verb, indikativ, preteritum, aktiv) till lemmat *klyva* samt 2) NCUSNI (substantiv, utrum, singular, nominativ, obestämd form) till lemmat *klöv*, medan vi i fallet *tår* möter analyserna 1) NCUPNI (substantiv, utrum, singular, nominativ, obestämd form) till lemmat *tår* samt 2) (substantiv, utrum, plural, nominativ, obestämd form) till lemmat *tå*. Vi har alltså här att göra med **extern homografi** (två eller flera lemmatiska former med identiska grafiska former men olika lemmabeteckningar). I ett senare steg används ett disambigueringsprogram som utifrån probabilistiska regler väljer rätt analys av de identiska grafiska formerna i löpande text (se Johansson Kokkinakis i denna volym).

Annoteringssystemet använder förutom SMDB också namnlistor (ca 250 000 namn i grundform samt genitivformer av dessa namn), listor över vanliga utländska ord samt listor över ca 300 förkortningar.Vi planerar också att lägga till de uppslagsord i NEO (Nationalencyklopediens ordbok, 1995) som saknar motsvarighet i SAOL samt vissa i ordlistan angivna former som inte kan genereras från uppslagsorden, t.ex. *sjöss* och *rättan* (dvs. reminiscenser från äldre språkskeden).

Som underlag för den här artikeln har vi valt att annotera Press97, en dagstidningskorpus i Språkbanken som omfattar ca 12 miljoner ord. Syftet är att få en bild av hur annoteringen kan användas för att sålla fram nyordskandidater till en kommande upplaga av SAOL. Resultatet av annoteringen visar att 90% av graforden i Press97 finns i SAOL, medan 8% av graforden utgörs av proprier och 0,2% av förkortningar. Återstoden är ord

som inte får någon analys, och det är dessa som diskuteras i följande avsnitt.

Att sålla bort skräp

Hur långt kan man automatisera processen med att ta fram förslag till nyord? Vid en snabb titt på det otaggade materialet kan man konstatera att en hel del arbete återstår när det gäller att eliminera ord som över huvud taget inte är intressanta. Framför allt gäller detta främmande ord. Dessa kan filtreras bort med hjälp av listor över vanliga ord i t.ex. engelska, franska, tyska och spanska. I synnerhet träffar man på formord och vanliga verb i engelskan: *the, of, at, you, said, was*. Vi möter också många förkortningar som saknas i förkortningslistorna. Dessa måste uppdateras efter hand eftersom vissa förkortningar bara förekommer inom specifika texttyper. Vanliga i tidningsspråk är till exempel *kd* (Kristdemokraterna) och *lib* (liberal). En del förkortningar förekommer i tabeller över sportresultat, t.ex. *mv* (målvakt) och *dug* (diskvalificerad under galopp). I bostadsannonser finns *nb* (nedre botten) och *rok* (rum och kök), och på matsidorna hittar man *msk* (matsked) och *krm* (kryddmått) m.m. En stor del av tidningstexten utgörs av egennamn (enligt NFO uppgår proprierna till 5% av graforden i sådan text.) Nya namn, i synnerhet personnamn, tillkommer kontinuerligt. Ortnamn däremot är relativt stabila. Namn på radio- och TV-program är högfrekventa i tidningstext, framförallt i de många radio- och TV-programtablåerna: *Ekonomiekot, Boktornet, Merhaba* och *Hundsjöviken*. Ett stort problem är att identifiera titlar på böcker, filmer och liknande. Dessa utgörs vanligen av längre syntagmer som t.ex. *Det perfekta vapnet* och *Jag är inte religiös*. För att spåra nya namn i löpande text behöver man ett namnigenkänningsprogram. Språkbanken har god kompetens inom detta område genom Dimitrios Kokkinakis (2001) och

ingår för närvarande också i ett nordiskt nätverk för namnigenkänning, Nomen Nescio.

En svårhanterlig grupp ord som inte lämpar sig för uppräkning i listform är referenser till olika beteckningssystem som t.ex. positionsangivelser (*57˚33’ N)* och referenser till schackfält och schackdrag *(e5, fxe5*).

Genom att använda en modul som analyserar okända sammansättningar bör man också kunna filtrera bort vissa produktiva sammansättningstyper som inte kommer att bli aktuella för en uppdatering av SAOL, t.ex. sammansättningar där proprier och akronymer urgör förled: *Cornelisrummet, Tokyobörsen, Göteborgsregionen, BNP-tillväxt, VM-kvalgrupp*.

Att sålla fram nyord

Den vanligt förekommande frågan om antalet ord i svenskan får ju ofta det svaret, att den uppgiften inte går att ge, eftersom framför allt möjligheterna till mer eller mindre självklara sammansättningar i princip är obegränsade. Det är därför ingen självklarhet hur fyllig en ordlista som SAOL skall vara. Förutom grundstommen i ordförrådet bör den förteckna också sådant som kanske inte dagligen möter i tidningstext (t.ex. vissa äldre ord, olika typer av fackord osv.), men den bör förstås också ge rikhaltiga exempel på sådana typer av nybildningar som speglar samtiden på skilda plan.

Med hjälp av SMDB får vi reda på att ca 20% av orden i SAOL12 inte finns belagda i den korpus på ca 60 miljoner ord ur Språkbanken som ligger till grund för frekvensframtagningen. Procentsatsen blir något högre om man nöjer sig med de 12 miljonerna i Press97, det material som vi tittar närmare på i denna uppsats.

Låt oss ta några exempel på sammansättningar (både förleds- och efterledssammansättningar) med ordet *affär*. SAOL förtecknar 56 ord som börjar på *affärs-* (*affärsangelägenhet, affärsanställd, affärsbank* etc.). 12 av dessa (t.ex. *affärsbegåv-*

ning, affärsfastighet och *affärsskylt*) saknar belägg i Press97, men i gengäld hittar man i samma material ytterligare 122 ord som **inte** finns med i SAOL (*affärsadvokat, affärsanalytiker*, *affärsansvar* etc.). Många av dessa ord verkar vara mer eller mindre tillfälliga bildningar skapade för just det aktuella sammanhanget. Andra klingar bättre och skulle kanske kunna platsa i en kommande upplaga av ordlistan, t.ex. *affärsinriktad*, *affärslunch* och *affärspartner.*

Ord i SAOL med *-affär* som efterled (t.ex. *speceriaffär*) finns medtagna till ett antal av 85, av vilka 12 saknas i Språkbanken som helhet och ytterligare 14 i just Press97. Som exempel kan nämnas *gottisaffär, skrädderiaffär* och *vitvaruaffär* (hela materialet) och *charkuteriaffär, kortvaru-affär* och *porslinsaffär* (Press97). Å andra sidan uppvisar Press97 ca 130 sammansättningar som SAOL inte förtecknar. Också här bär många tillfälligheten prägel, men ett och annat av orden kan säkert kandidera till en plats i kommande upplaga av SAOL, t.ex. *andrahandsaffär* (som komplement till *second hand-affär*), *företagsaffär, skalbolagsaffär* och *storaffär*. Att antalet efterledsammansättningar i korpusen blir så stort beror på det starkt varierande semantiska innehållet i *affär*, något som inte framgår av ordlistan (där ordet saknar betydelseangivelse) men som aktualiseras i de många skiftande typerna av sammansättningar (jfr ordlistans *blomsteraffär*, *insideraffär, kärleksaffär* etc.)

I bedömningen av presumtiva kandidater till sammansättningar i ordlistan spelar givetvis frekvensen på ordet och antalet använda böjningsformer en stor roll. Vid en huvudsakligen manuellt genomförd lemmatisering av ca 25% av det otaggade materialet i Press97 ställde vi kravet på minst tre böjningsformer samt minst två belägg på respektive böjningsform. Ett 70-tal lemman blev resultatet. Inte oväntat verkar många av dessa vara självförklarande nog för att inte behöva tillföras en ordlista av SAOL:s snitt, men åtskilliga mer eller mindre självklara nyordskandidater kan noteras, bland substantiven *analysgrupp, antirasist, bankfusion, bilstol, lagkamrat, landslagsman, lyxkrog, länklista, lönearbete,*

valfråga, valobservatör, videokonferens, videoskärm, vårdkö, vänsterpartist och *världscup,* bland adjektiven *användarvänlig, arbetsrättslig, avskalad* och *verklighetsbaserad* samt verbet *ljussätta.*

Möjligheter – i dag och i morgon

Med hjälp av SMDB kan vi i dag komma åt varje uppslagsord i SAOL och samtidigt få frekvenser för såväl grundform som böjningsformer, frekvenser baserade på Språkbankens nuspråkliga korpusar (ca 60 miljoner ord). Så utnyttjad fungerar alltså databasen som en lemmatiserare och frekvensräknare. Söker man på ordet *gata* dyker substantivets samtliga åtta böjningsformer upp med sina respektive frekvenser (*gata* 1070, *gatas* 5, *gator* 3367 etc.). Begär man ordet *kollega* får man dessutom klart för sig frekvensförhållandena mellan de två alternativa pluralformerna *kolleger* (1571) och *kollegor* (751). SMDB kan alltså ge snabba besked om språkbruket, vilket på många punkter kommer att vara till stor hjälp i arbetet med SAOL13. Inte minst gäller detta en uppstramning av böjningsuppgifterna vid de mängder av ord som i dag anges ha alternativa böjningar. Vid *hasch* och *morfin* (i SMDB försedda med gruppbeteckningen 4h för ord med varierande genus utan pluralformer) har SAOL12 *-en* el. *-et* resp. *-et* el. *-en*. SMDB ger oss uppgifter som bekräftar vår intuition, nämligen att neutrumformen bör anges som den primära, medan utrumformen kan reduceras till en biform.

När hela Språkbanken väl annoterats, kommer det också att dyka upp böjningsformer som inte förutsågs vid gruppindelningen av SAOL-materialet. Orden *ambivalens, lekfullhet, lyssning* och *vänskap* tillfördes vid klassificeringen grupp 34 (utrala ord utan pluralbildning), och följaktligen slinker de i Press97 använda formerna *ambivalenser, lekfullheter, lyssningar* och *vänskaper* igenom fullformsnätet. Vi får alltså via texterna ett facit på uppblåsningen och kan i vissa fall komma

att behöva göra ändringar i vår indelning, ändringar som också kommer att ge nedslag i kommande tryckta upplagor av SAOL.

Korrektheten vid lemmatiseringen i SMDB begränsas än så länge av homografin. Frekvensen för *barn* omfattar alltså både singular och plural, frekvensen för *koppar* gäller både pluralformen av *kopp* och metallen i obestämd form. Till synes helt korrekta serier kan också innehålla betydande felaktigheter av andra skäl. Begär man verbet *duga* får man sifferuppgifter för infinitiv, presens, preteritum, particip och imperativ. Man gör den iakttagelsen att variantformen *dugde* med sina två belägg lever farligt inför nästa upplaga i förhållande till den starka formen *dög* (272 belägg), men man konstaterar också med stor förvåning siffran 235 för den från semantisk synpunkt märkliga imperativen *dug!*. En närmare undersökning av beläggen visar också att det i samtliga fall rör sig om en i sportartiklar vanlig förkortning för ’diskvalificerad under galopp’ (se ovan). Homografifloran – rik nog som den redan är i svenskan – kan alltså få näring från de mest oväntade håll. Det skall bli verkligt spännande att så småningom ta del av resultaten av det ovan nämnda disambigueringsprogrammet, som skall analysera de identiskt grafiska formerna i löpande text.

Avslutning

Martin! Du har ägnat mycket av din aktiva forskartid åt uppbyggnaden av Språkbanken och tar också kontinuerligt och konstruktivt del av det arbete som vi velat presentera i denna uppsats. Vi tycker oss ha kommit en bra bit på väg mot en välstrukturerad och innehållsdiger svensk morfologisk databas, som bland mycket annat kan leda till nya förbättrade versioner av såväl SAOL som andra nuspråkliga ordböcker. Välkommen att fortsätta samarbetet med oss också efter passerat pensionsdatum!

Referenser

Allén, S. 1970. *Nusvensk frekvensordbok. 1. Graford, homografkomponenter*. Data linguistica 1. Göteborgs universitet.

Allén, S. 1971. *Nusvensk frekvensordbok. 2. Lemman.* Data linguistica 4. Göteborgs universitet.

Berg, S. 1978. *Olika lika ord. Svenskt homograflexikon.* Data linguistica 12. Göteborgs universitet.

Berg, S. & K. Samuelson 1988. SAOL as a Spelling-Checker Dictionary. I: *Studies in Computer-Aided Lexicology*. Data Linguistica 18. Göteborgs universitet.

Berg, S., Y. Cederholm & M. Gellerstam 2001. Svensk morfologisk databas baserad på tolfte upplagan av Svenska Akademiens ordlista. (Under arbete.)

Danielsson, P. & J. Järborg 1996. *Morphosyntactic Description of Swedish*. PAROLE report (WP-4.2.2b).

EAGLES *Recommendations for the Morphosyntactic Annotation of Corpora*. 1996. EAGLES Document EAG-TCWG-MAC/R.

Ejerhed, E., G. Källgren, O. Wennstedt & M. Åström 1992. *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project.* DGL-UUM-R-33. Department of General Linguistics, University of Umeå. Report No. 33.

Gellerstam, M., Y. Cederholm & T. Rasmark 2000. The Bank of Swedish. *Proceedings of the 2:nd Language Resources and Evaluation Conference (LREC),* Athens. (Samt i: Research Reports from the Department of Swedish, Göteborg University. GU-ISS-00-6).

Kokkinakis, D. 2001. *Design, Implementation and Evaluation of a Named-Entity Recognizer for Swedish.* Research Reports from the Department of Swedish, Göteborg University. GU-ISS-01-1.

Kokkinakis, D. & S. Johansson Kokkinakis 1997. *A Robust and Modularized Lemmatizer/Tagger for Swedish Based on Large Lexical Resources.* Research Reports from the Department of Swedish, Göteborg University. GU-ISS-97-1.

NEO = *Nationalencyklopedins ordbok* 1–3. 1995–96. Höganäs: Bra böcker.

Nomen Nescio. Nordiskt nätverk för namnigenkänning. <http://spraakbanken.gu.se/nn/>

PAROLE. Institutionen för svenska språket. <http://spraakdata.gu.se/lb/parole/>

SAOL12 = *Svenska Akademiens ordlista över svenska språket.* 12 uppl. 1998. Stockholm: Norstedts.