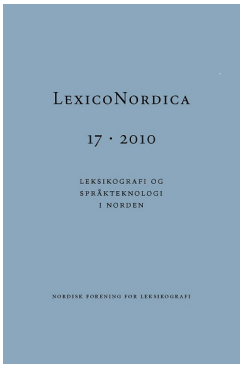
LexicoNordica

Titel: Med Zipf mot framtiden - integrerad lexikonresurs för svensk språkteknologi

Forfatter: Lars Borin

Kilde: LexicoNordica 17, 2010, s. 35-54

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



Søgbarhed

Artiklerne i de ældre LexicoNordica (1-16) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

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.

---

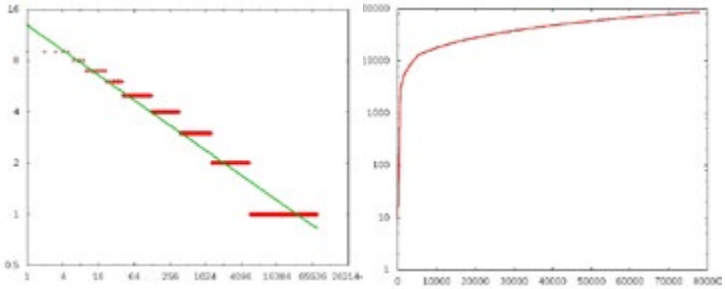
inarbetade engelska termen *lemma* 'grundform' visat sig leda till oönskad begreppsförvirring när man försöker använda samma term för denna formellt definierade enhet (som till råga på allt ligger nära det som på engelska brukar kallas *lexeme*). Därför har vi valt en helt ny term för den; det vi här 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).

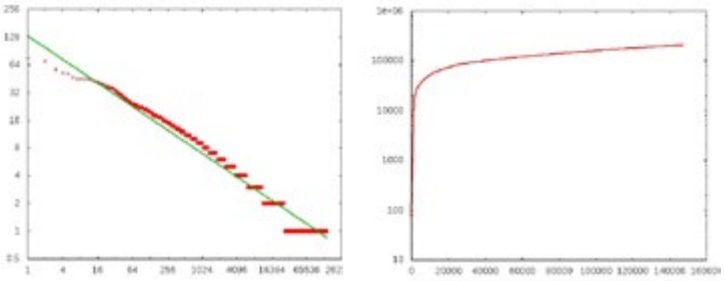


a: rangordning – antal betydelser b: rangordning – betydelsetillväxt

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