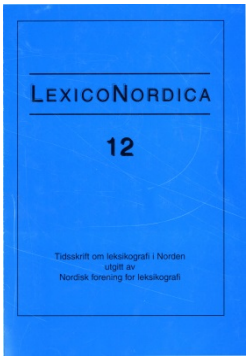
LexicoNordica

Titel:	Mannen är faderns mormor: *Svenskt associationslexikon* reinkarnerat
Forfatter:	Lars Borin
Kilde:	LexicoNordica 12, 2005, s. 39-55
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.

Lars Borin

Mannen är faderns mormor: *Svenskt associationslexikon* reinkarnerat

Svenskt associationslexikon (SAL; Lönngren 1992) is relatively new and relatively little known Swedish thesaurus, compiled on the basis of corpora and some existing Swedish monolingual lexical resources. SAL is organized as a strict lexical-semantic hierarchy originating in an artificial top lexeme, where (primary) vertical 'parent'–'child' relations point to less central, but semantically closely related lexemes, and (secondary) horizontal 'sibling' relations form thesaurus-like word families. This paper describes my work – in collaboration with the author of SAL, Lennart Lönngren – on making an electronic version of the full SAL (comprising 71752 entries) publicly available through Språkbanken at Göteborg University in a modern, standardized format, which has been thoroughly checked for formal errors, enhanced with additional information about lemmas and inflectional paradigms of entries, and made browsable with a web-based graphical interface capable of displaying and navigating the network structure of SAL.

1. SAL – Svenskt associationslexikon

Svenskt associationslexikon (SAL) är en relativt ny och relativt omfattande svensk tesaurus som dock är föga känd och därmed också lite använd. SAL skapades under åren 1987–1992 under ledning av Lennart Lönngren, då verksam vid Centrum för datorlingvistik och Slaviska institutionen, båda vid Uppsala universitet.[1] Den har givits ut enbart i två små stencilupplagor i form av rapporter från Centrum för datorlingvistik, Uppsala universitet (Lönngren 1988), samt Institutionen för lingvistik, Uppsala universitet (Lönngren 1992). Dessutom har ända från början uppslagsorden och deras mest grundläggande semantiska relationer (se nedan) förelegat i elektronisk form, som rena textfiler.

SAL:s tillkomsthistoria dokumenteras i Lönngren 1989. De första källorna var korpusar: en lärobok i svenska för invandrare och ett populärvetenskapligt textmaterial. Dessutom innehåller SAL en relativt stor mängd – ca 3000 – egennamn från olika källor, framförallt en liten encyklopedi (3215 ingångar inleds med versal i den elektroniska ver-

[1] Lennart Lönngren är sedan 1993 verksam vid Ryska institutionen, Universitetet i Tromsø (se <http://www.hum.uit.no/a/lonngren/>). Andra medverkande i arbetet med SAL var Gunilla Fredriksson som arbetade med lexikondefinitionerna samt Ágnes Kilár som svarade för programmeringen i projektet.

sionen av SAL, men bland dem återfinns också sådana uppslagsord som *A-vitamin*). Så småningom utökades ordförrådet betydligt med hjälp av en lista av stickord ur Svensk ordbok (1986) som införskaffades från Språkdata, Göteborgs universitet. Den andra pappersversionen av SAL (Lönngren 1992) innehöll 71.750 lexikoningångar.

SAL är som sagt ett slags tesaurus, som är organiserad enligt följande principer eller hypoteser:

(1) **associationsprincipen**; ”sett från ett visst lexem kan det övriga ordförrådet uppdelas i semantiskt relaterade och orelaterade lexem. Relaterade lexem kan vi kalla semantiska grannar.” (Lönngren 1989:7) Här kan man vara mer exakt, och tala om semantiskt direkt relaterade, indirekt relaterade och orelaterade lexem. De semantiska grannarna blir då den första kategorin. Den andra kategorin innehåller grannars grannar, grannars grannars grannar, etc., och den tredje kategorin innehåller sådana lexem som inte kan nås via en kedja av semantiska grannar i ett visst antal steg. Lönngrens ursprungliga definition innehåller dessa tre kategorier implicit (åtminstone har jag antagit att användandet av ordet ”grannar” implicerar att man ska göra en skillnad mellan direkta och indirekta semantiska relationer), så här explicitgör jag dem bara.

(2) **centralitetspricipen**; ett språks ordförråd kan ordnas i mer eller mindre centrala ord.

(3) **lexemprincipen**; lexikoningångarna är semantiska enheter och i princip spelar formella faktorer ingen roll för att avgöra hur enheterna avgränsas inbördes och vilka enheter som postuleras. Sålunda innehåller SAL inga ordklassangivelser: ”För vårt vidkommande är det t.ex. helt oväsentligt om ordet *alternativ* används som substantiv eller adjektiv” (Lönngren 1989:17). Fast helt oväsentliga är ordklasserna tydligen inte ändå: ”dock brukar vi skilja verb och substantiv åt, t.ex. *resa* delas upp i två enheter” (Lönngren 1989:17). I enlighet med den här principen spelar ordklassen heller ingen roll för de direkta lexikaliska relationerna, utan de rör sig ledigt över ordklassgränserna, som vi kommer att se längre fram.

(4) **anti-isolationismprincipen**; lexikonet ska inte innehålla ”isolerade” lexem. Alla lexem ska ha en eller flera semantiska relationer till andra lexem i lexikonet. På det viset bildas ett antal ”kontinenter” med mer eller mindre relaterade lexem. Eventuellt är det till och med så att varje lexem i lexikonet är

nåbart från varje annat lexem utan att man behöver gå via det artificiella topplexemet PRIM (se nedan), men det återstår att kontrollera.

(5) **hierarkiprincipen**; lexikonets ”ryggrad” ska vara hierarkisk. Den hierarkiska organisationen ansluter till princip 1 och 2 ovan, på det viset att nästan alla lexem i lexikonet ges en sorts ”förklaring” i form av en centralare (princip 2) semantisk granne (princip 1). I många fall är det också ändamålsenligt att ange ytterligare en semantisk granne för att ”förfina” förklaringen. Hierarkin ska vara strikt, en trädstruktur och inte en allmän grafstruktur. Det får inte finnas några cirklar i den med andra ord; ett lexem får inte direkt eller indirekt förklaras med hjälp av sig själv.

Lönngren använder några olika namn om de förklarande lexemen, även om det alltid är fråga om samma relationer. Emellanåt kallas de förklarande lexemen ”deskriptorer”, där den första, obligatoriska semantiska grannen blir ”huvuddeskriptor” och den andra ”determinerande deskriptor”. Speciellt när deskriptorerna, eller de primitiva (ohärledda) relationerna, diskuteras tillsammans med ett antal härledda (men fortfarande direkta) relationer, så kallas huvuddeskriptorn ”moder” och den determinerande deskriptorn kallas för ”fader”. Här ger jag några exempel på lexem och deras deskriptorer, på måfå hämtade från bokstaven B i Lönngren 1992:

bränsletillägg: tilläggskostnad + bränsle
brödföda: uppehälle
bröllop: gifta sig
bröstbarn: suga + mjölk
Bulgakov: författare + rysk

Varje lexem i SAL ska alltså ha en moder och optionellt en fader. För att uppfylla detta villkor och samtidigt undvika cirkularitet i en ändlig ordmängd, införs ett artificiellt ”topplexem”, kallat PRIM. Lexikonets centralaste lexem, som själva alltså logiskt inte kan ha något centralare förklarande lexem, får sålunda PRIM som (huvud)deskriptor/moder. I det ursprungliga SAL fanns det 54 sådana lexem. Ur mödra- och fadersrelationen härleds sedan en rad andra direkta ”grannrelationer”, som jag återkommer till nedan. Det är framför allt dessa härledda relationer som gör SAL till ett slags tesaurus, eftersom ord med gemensamma föräldrar tenderar att flockas i semantiska grupper. Samtidigt skiljer sig också SAL från en typisk tesaurus framför allt genom att alla ordklasser finns i SAL, även de slutna, grammatiska klasserna samt

egennamn. Pronomen, prepositioner och proprier är fullvärdiga SAL-lexem, medan sådana brukar lysa med sin frånvaro i vanliga tesaurusar.

2. Mot en ny version av SAL

Jag arbetade under många år på Centrum för datorlingvistik och sedermera, efter en institutionssammanslagning, på Institutionen för lingvistik vid Uppsala universitet, där jag kunde följa arbetet med SAL på nära håll när det pågick, även om jag själv inte var inblandad i det. Med tiden har jag förstått att denna lexikala resurs är nästintill okänd,[2] vilket är synd, därför att den borde kunna vara mycket användbar både för språkvetenskapliga och språkteknologiska ändamål om den vore mer allmänt känd. Detta fordrar naturligtvis att den är mer allmänt tillgänglig, t.ex. genom Språkbanken <http://spraakbanken.gu.se> vid Göteborgs universitet.

Ungefär ett år efter det att jag hade kommit till Språkdata och Språkbanken, i november 2003, tog jag kontakt med Lennart Lönngren och frågade om han fortfarande hade kvar det digitala underlaget för SAL, och även vad han skulle tycka om att göra lexikonet allmänt tillgängligt i Språkbanken för uppslagning via ett webbgränssnitt samt i dess helhet i elektronisk version för språkteknologiforskare. Han ställde sig mycket positiv till detta och skickade mig strax SAL i form av fyra textfiler med sammanlagt 71.750 rader/lexikoningångar. Raderna är strukturerade på ett av följande två sätt:

```
_&lexem&deskriptor&&
_&lexem&huvuddeskriptor&&determinerande
deskriptor&
```

Ytterligare ett år senare, i oktober 2004, satte jag igång med att förbereda en språkbanksversion av SAL. Detta arbete har hittills bestått i fyra rätt olika, men sammanlänkade, aktiviteter:

(1) Formell kontroll av lexikonstruktur och transformation av lexikoninformationen till det format som ska användas i Språkbanken. Detta gjordes med en uppsättning datorprogram skrivna av mig.
(2) Övervägande av vilket lexikonformat som ska användas i Språkbanken och vilken information som det vidareutvecklade SAL ska innehålla, liksom definition av ett s.k. API

[2] En Google-sökning efter ”svenskt associationslexikon” ger två träffar, och motsvarande sökning på engelska (”swedish associative thesaurus”) ger också två relevanta träffar.

(*Application Program Interface*; ett gränssnitt för datorprogram som ska använda sig av informationen i SAL), med beaktande av relevanta liknande arbeten på andra håll i världen med språkteknologiska lexikonmaterial.

(3) Införande av ytterligare information i SAL. Detta gjordes delvis manuellt, delvis automatiskt med datorprogram skrivna av mig.

(4) Utvecklande av ett webbaserat grafiskt gränssnitt för användare som vill slå i SAL. Gränssnittet utvecklas just nu som ett datalingvistiskt examensprojektarbete av Isabelle Cabrera. Det har både textuella och mer rent grafiska komponenter. I de förra presenteras lexikoningångar mer eller mindre som i pappersversionen av SAL (se figur 1 nedan), och i de senare återfinns såväl trädstrukturer för navigering i lexikonet ungefär som i en dators filsystem, som en visualisering av lexem som noder i ett slags ”semantiskt nätverk” (se figur 2 nedan).

De följande avsnitten innehåller en lite utförligare beskrivning av var och en av aktiviteterna (1) – (4).

2.1. Formell kontroll och transformation av lexikoninformation

En av de första åtgärderna som behövde vidtas för att bereda SAL för språkbanksbruk var en formell kontroll av den lexikaliska hierarkin. Detta gjordes med ett par särskilt för ändamålet skrivna datorprogram. Först kontrollerades hierarkin på mödernet. Det visade sig då att några lexem i själva verket hade en cirkulär förklaring. Sammanlagt ingick 574 lexem i en eller annan cirkulär struktur, varav lejonparten orsakades av en direkt cirkulär förklaring högt uppe i hierarkin, som berörde 492 lexem:

stänga: öppen + inte
öppen: stänga + inte

Lennart Lönngren ställde villigt upp och rättade till de aktuella lexikoningångarna. Som ett resultat av detta fick lexikonet ett nytt lexem och PRIM fick ytterligare ett barn (*öppen*). Nästa kontroll handlade om huruvida det fanns några cirkulära strukturer på fädernet. På samma sätt som i fallet med mödrarelationerna visade det sig finnas några faderscirklar. Dessa berörde ett litet antal lexem, inalles 14 stycken.

Igen rätade Lennart Lönngren ut cirklarna, och lade till ytterligare ett lexem som ett resultat av detta. Efter dessa formella kontroller och de rättelser och tillägg som de resulterade i, kom SAL att innehålla 71752 lexem och PRIM kom att ha 55 barn (se tabell 1).

Tabell 1: *PRIM-barnen – de 55 mest centrala lexemen i SAL*

all	göra	mot[1]	tal[1]
annan	ha[1]	mycken	till
använda	hur	måste	tänka
att	hända	namn	vad[1]
bara	i[2]	natur[1]	var[1]
bra[1]	inte	när[1]	vara[1]
den[1]	ja	och	varm
fort[1]	just[1]	om[1]	vem
framme	kant	om[2]	veta
färg[1]	kunna	på	vid[1]
för[2]	ljud	rak	vilja[1]
förbi[1]	ljus[1]	röra[1]	än[1]
före[1]	med[1]	så[1]	öppen
genom[1]	men[1]	säga[1]	

Sedan vidtog bearbetningar för att explicitgöra implicit information i SAL. En sorts information som finns i pappersversionen av SAL men inte i det grundläggande digitala formatet är de härledda relationerna. Lönngren (1989) räknar med 9 härledda relationer, som måste anses vara direkta semantiska relationer i de termer som diskuterades i avsnitt 1 (det innebär att de inblandade lexemen är ”semantiskt direkt relaterade”).[3] Här utvidgas samma familjemetafor som gav upphov till benämningarna moder och fader. Lönngren (1988, 1989, 1992, 1998) använder ett slags sifferkoder för både primitiva och härledda relationer (se tabell 2 på nästa sida).

[3] En viktig anledning till att kalla även de härledda relationerna för direkta relationer har att göra med de primitiva relationernas tillkomstprocess, där vissa logiskt senare härledda relationer i själva verket visar sig vara faktiskt tidigare, eller åtminstone samtidiga, och kanske egentligen ska betraktas som primärare: ”Under arbetets gång har välformade syskongrupper alltmer klart framstått som det viktigaste korrektivet när det gäller att utse föräldrar till ett lexem. [...] Man kan faktiskt se hela lexikonet som ett försök att skapa optimala horisontella serier av ord; de vertikala förbindelserna ger enligt min uppfattning inte lika starkt intryck av associativ samhörighet och bör därmed komma i andra hand.” (Lönngren 1989:11)

I figur 1 på nästa sida visas två lexikoningångar ur SAL (Lönngren 1992) som får tjäna som konkret illustration av några av dessa relationer.

TABELL 2: *Relationer (”deskriptorer”) i SAL*

1– ”jag är barn till dig”	2– ”jag är förälder till dig”
11 ”du är mor” 12 ”du är far”	21 ”jag är mor” 22 ”jag är far”
3– ”jag är gift med dig”	4– ”jag är syskon till dig”
31 ”jag är maka” 32 ”jag är make”	40 ”vi har far och mor gemensamma” 41 ”vi har gemensam mor” 42 ”vi har gemensam far” 43 ”min mor är din far” 44 ”min far är din mor”

unge[1]: 11 djur; 12 ung; 21 busunge, kull[1], yngel; 22 ankunge, däggdjursunge, ejderunge, finkunge, fågelunge, föl, gässling, kalv, kattunge, killing, kyckling, lamm, ormyngel, spädgris, sälunge, ung-stadium, valp, vivipar, älling; 31 busig; 32 and[1], anka[1], däggdjur, ejder, fink, fågel, får, föda[1], get, gris, gås, hund, häst[1], hönsfågel, katt[1], ko, orm, stadium, säl; 41 ... (djur 21); 42 barn... (ung 22); 43 ... (djur 22); 44 benjamin, föryngra, föryngring... (ung 21)

unge[2]: 11 barnunge; 21 rännstensunge, satunge, snorunge, trasunge; 31 busig, gata, trasa[1]; 41 glytt

FIGUR 1: *Två lexikoningångar i SAL*

Relationskoderna är i viss mån inbördes motiverade, men de har egentligen inte en riktig formell intern struktur – de är inte strikt kompositionella – utan strukturen är ”sluten” i den meningen att man inte kan se hur man skulle kunna skapa koder för andra möjliga härledda relationer, samtidigt som den möjligheten förutses:[4] ”Om man vill kan man

[4] Därför har jag i den nya elektroniska versionen av SAL ersatt Lönngrens släktskapskoder med nya, strikt kompositionella sådana. Samma härledda relationer används som i det ursprungliga SAL, men nu finns också en reell möjlighet att

naturligtvis fortsätta och finna mormor och morfar, farmor och farfar, samt barnbarn.” (Lönngren 1989:15) En konkret mormorsrelation i SAL har inspirerat till titeln på denna artikel:[5]

fader: far
far: man[2] + familj

2.2. Språkbankens lexikonformat och SAL

2.2.1. Bakgrund

Under senare år har vi sett en utveckling inom språkteknologi och andra relevanta informationsteknologiska delområden mot större integration av resurser och system. Denna utveckling har flera orsaker, men två av de viktigaste är förmodligen å ena sidan en trend inom språkteknologi som innebär att man arbetar med allt större resurser som förses med alltmer sofistikerade annotationer, och å den andra det alltmer omfattande hopknytandet av den informationsteknologiska världen genom framförallt WWW. Eftersom uppbyggande och annotering av stora språkteknologiska resurser är mycket arbetskrävande och därmed mycket dyrt, finns det en stark önskan att kunna använda existerande resurser i så många sammanhang som möjligt. Därför ser vi idag en koordinering av språkteknologistandardiseringsansatser i ISO:s regi, i form av TC37/SC04 <http://www.tc37sc4.org>. Samtidigt pågår ett idogt arbete i WWW-världen – samordnat av *World Wide Web Consortium* (W3C; <http://www.w3.org>) – vars mål är att få fram format och verktyg som ska göra det möjligt att leta efter resurser på webben genom att man anger vad för slags innehåll man är intresserad av, snarare än i vilka ord innehållet råkar uttryckas, samt göra det möjligt för maskiner att kombinera information från olika källor för att på det viset skapa den önskade informationen. Målet med detta arbete brukar karakteriseras med uttrycket *semantiska webben* (Semantic Web).

Under denna rubrik finner man egentligen två sorters aktiviteter. Dels arbetar man med begreppshierarkier med åtföljande kontrollerade vokabulärer, som i de här kretsarna brukar kallas för *ontologier* (som ska kunna användas för att märka upp webbresurser med strukturerad information om resurserna ifråga), dels utvecklar man också generella representations-, lagrings- och överföringsformat för information ut-

---

införa nya relationer med känd betydelse genom att man kombinerar befintliga relationskoder. Här är dock inte platsen att gå in på detaljerna i de nya koderna.

[5] Titen alluderar förstås också till Wordsworths aforism ”Barnet är mannens fader”.

tryckt i sådana kontrollerade vokabulärer, samt formalismer och verktyg för att utföra formellt sund slutsatsdragning över data i dessa format, bland annat i syfte att kunna jämka ihop information från många informationskällor och i samband med detta explicitgöra implicit information.

I själva verket kompletterar de två aktiviteterna – arbetet med språkteknologistandarder och det med semantiska webben – varandra ganska väl; arbetet med ontologier för semantiska webben bygger med nödvändighet i mångt och mycket på traditionell terminologisk och i viss mån lexikologisk kunskap, medan de kunskapsrepresentationsformat som växer fram för den semantiska webben i sin tur mycket väl bör kunna användas också för att representera den sorts språkliga kunskap som är hjärtat i språkteknologiska tillämpningar. Detta har man börjat ta fasta på på många håll, både för allmänna språkteknologiska tillämpningar (t.ex. Bontcheva et al. 2003), men även mer specifikt för språkteknologiska lexikonresurser (Ide et al. 2003).

I Språkbanken arbetar vi sedan en tid tillbaka med att anpassa format och verktyg för att de ska vara kompatibla med denna utveckling. Arbetet i Språkbanken handlar om alla aspekter av språkteknologiska resurser, men med ett fokus på korpusar och korpusannotationer samt lexikaliska resurser. Precis som Språkbankens övriga resurser har lexikonresurserna två typer av användning: som språkliga resurser – alltså för mänskligt bruk – och som språkteknologiska resurser, dvs. som informationskälla i språkteknologiska applikationer.

Vi experimenterar f.n. med ett dataformat som kallas RDF (Resource Description Framework) som bärare av lexikalisk och annan språklig information, och det är det formatet som den senaste versionen av SAL föreligger i just nu. RDF är den grundläggande byggstenen i W3C:s olika format för den semantiska webben. Här är inte den rätta platsen att gå in på detaljer, utan den intresserade hänvisas till W3C:s webbplats, där det finns mängder av information om RDF och andra format som bygger på det. Att notera här är bara att om man vill arbeta med RDF som representationsspråk så får man automatiskt tillgång till en stor och snabbt växande mängd verktyg för att göra detta. För SAL har vi valt att arbeta med ett fritt tillgängligt RDF-databassystem med öppen källkod som heter Sesame (Broekstra 2005; se även <http://www.openrdf.org>).

2.2.2. Ny information i SAL

Under arbetet med SAL har det blivit aktuellt att lägga till information av olika slag. Till en del handlar det om information som är härledbar ur grundformatet utan att man behöver någon ytterligare kunskapskälla, alltså samma sorts implicita information som de härledda relationerna som beskrevs i avsnitt 2.1. F.n. finns en sorts sådan tilläggsinformation i SAL-databasen, nämligen information om lexemens ”lexikaliska djup”. Eftersom SAL är en strikt hierarki som avslutas i PRIM, så kan varje lexem tilldelas ett ”djup” eller en ”nivå”, nämligen hur långt ifrån PRIM lexemet ligger om man följer en väg som går enbart via lexikaliska mödrar. Det visar sig att det djupast liggande – eller minst centrala – lexemet i SAL, *okynnig*, får nivå 15 (se tabell 4).

På samma sätt kunde man tänka sig att ge lexemen en nivå på fädernet, men här blir det genast besvärligare. Först och främst så är fadern optionell, så vi kan inte räkna med att det ska finnas en obruten kedja upp till PRIM (ifall man inte räknar en saknad fader som att PRIM är fader). Det finns dessutom inget krav på att fadern ska vara centralare än barnet; den ska ju huvudsakligen differentiera förklaringar där samma moder används. Som en slags kompromiss har jag tilldelat varje lexem ett fadersdjup, som är 0 (noll) ifall fader saknas och annars faderlexemets lexikala djup enligt ovan. Största fadersdjup i SAL uppvisar lexemet *tjuvpojksglimt*, vars fader är *okynnig* med det lexikala djupet 15 (men dess eget lexikala djup är bara 4: *tjuvpojksglimt* < *blick* < *se* < *ljus*[1] < PRIM).

TABELL 4: *Det minst centrala lexemet i SAL*

0	PRIM
1	< göra
2	< arbeta
3	< möda
4	< ansträngа
5	< svår
6	< besvärlig
7	< krångla
8	< bråk
9	< bråka
10	< busa
11	< rackare
12	< rackartyg
13	< ofog
14	< okynne
15	< okynnig

För att SAL ska bli riktigt användbar som språkteknologiresurs behöver man också tillföra en del ny information utifrån, främst av formell karaktär. Som tidigare nämnts innehåller SAL inte ens ordklassangivelser, och tabell 4 illustrerar på ett utmärkt sätt hur mödrarelationen i SAL – och därmed förstås också härledda relationer – ignorerar ordklassgränserna; i tabell 4 återfinns substantiv, verb och adjektiv om vartannat. En resurs som SAL blir mångdubbelt värdefullare för språkteknologiska ändamål ifall man i tillägg till den lexikalisk-semantiska och encyklopediska information som redan finns där i form av de associativa relationerna också kan få information om formella språkliga egenskaper som t.ex. ordklass, böjning och syntaktisk valens.

Av den anledningen har jag lagt till information om lemman och börjat lägga till information om böjningsmönster i SAL-databasen. Rent praktiskt har detta skett så att SAL har halvautomatiskt samkörts med NEO-databasen, den databas som legat till grund för Nationalencyklopedins ordbok (1995).[6] Ur NEO-databasen har sedan tillhörande lemman med ordklassangivelser och böjningsinformation hämtats för de lexem som kunnat identifieras genom samkörningen. För övriga lexem har lemman och information om ordklass lagts till, men ännu ingen böjning.[7] Böjningsinformationen kommer så småningom att ligga till grund för en morfologisk analysmodul baserad på s.k. funktionell morfologi (Forsberg & Ranta 2004). I SAL-databasen finns således för närvarande följande information, allt i RDF-format i en Sesamedatabas:

- unikt identifierade SAL-lexem[8]

---

[6] Ungefär 48 000 SAL-lexem hade endast en motsvarighet i NEO-databasen, och den har antagits vara den rätta utan manuell kontroll. Eftersom NEO är efterföljaren till Svensk ordbok som ju har tillhandahållit en stor del av ordförrådet i SAL är detta antagande knappast särskilt farligt. Knappt 12 500 SAL-lexem finns inte alls i NEO-databasen. De sistnämnda utgör en blandad kompott; här finns en hel del egennamn och flerordenheter, men även sammansättningar och avledningar som visserligen ibland finns med i NEO, men inte som egna ingångar, samt en hel del främst naturvetenskapliga fackord ur det populärvetenskapliga korpusmaterialet. Dessa, liksom de resterande drygt 11 000 lexemen (där det fanns flera lexem med samma uppslagsform åtminstone i en av SAL eller NEO) har behandlats manuellt.

[7] Ordklassuppsättningen har utökats med särskilda klasser för flerordingar – pmm ’flerordsegennamn’, vbm ’flerordsverb’, etc. – och förkortningar – nna ’förkortning, substantiv’, aba ’förkortning, adverb’, etc. Exakt hur böjningsinformationen ska utformas för flerordingarna är fortfarande en öppen fråga.

[8] Att en enhet i databasen är ”unikt identifierad” innebär att den har ett ID som identifierar just den enheten och ingen annan, som pekar ut just denna ”lingvistiska individ”. Detta är nyckeln till att koppla ihop olika språkliga resurser med varandra.

- SAL-relationer mellan SAL-lexem (11 olika relationer)
- lexikaliskt djup och lexikaliskt djup för eventuell fader
- grundform
- koppling till motsvarande NEO-lexem, unikt identifierade
- koppling till lemman, unikt identifierade, med information om ordklass och (än så länge icke-formaliserad information om) böjningsmönster

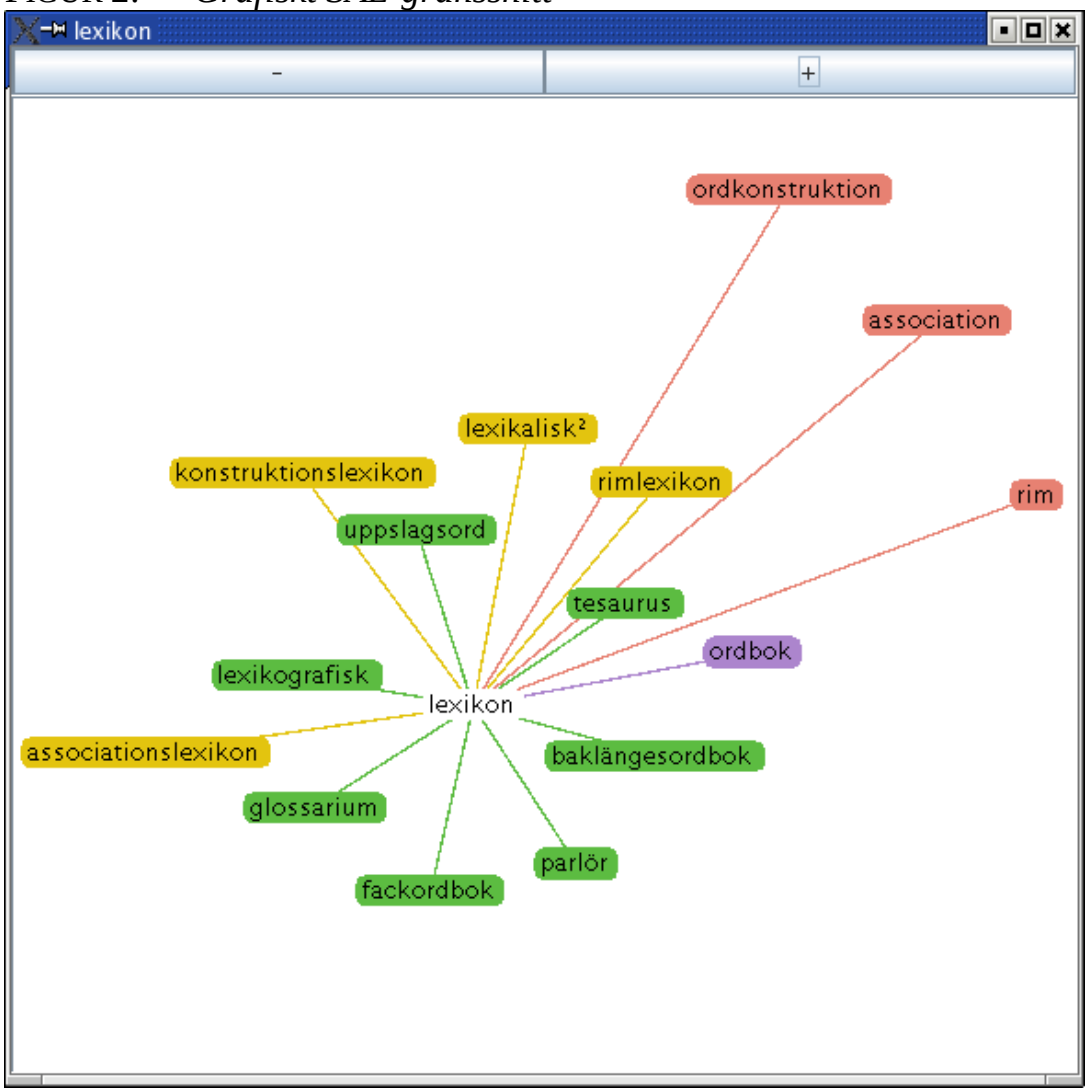
2.2.3. Användargränssnittet

Sesame-databassystemet innehåller redan ett webbaserat gränssnitt för att söka i databaserna, men det är inte på något sätt speciellt anpassat till lexikaliska eller ens språkliga data, utan kräver ganska mycket av sin användare i fråga om kunskaper om både RDF och databaser. Även om vi anser att RDF är ett bra underliggande representationsformat av skäl som jag redan har varit inne på, följer inte därav att det också skulle vara ett bra format för slutanvändaren att utsättas för. I själva verket har RDF skapats för maskiner och inte för människor. Därför behövs alltid ett skyddande lager – ett användargränssnitt – mellan RDF-formatet och användarna.

Sedan en tid pågår arbetet med att skapa en prototyp till ett sådant gränssnitt för SAL, som samtidigt ska kunna användas även för andra lexikonresurser samt utan problem kunna samverka med t.ex. Språkbankens korpussöksystem.

En första fungerande version av gränssnittet föreligger, där man kan slå upp lexem genom att söka i ett index eller genom att skriva en grundform eller en partiell grundform i ett textfält. Om det sistnämnda ger mer än en lexikoningång får man välja bland alternativen i en lista. När man väl har valt ett lexem, visas det på två sätt, (1) i form av en graf där noders och länkars färger talar om vilken relation det handlar om, och avståndet mellan noder visar hur nära de är relaterade, och (2) i en textruta med en utformning som efterliknar ingångarna i pappersversionen av SAL, fast utan de förkortningskonventioner som användes där av utrymmesskäl (tre punkter tillsammans med en hänvisning till en annan lexikoningång; se figur 1 ovan). Alla orden i såväl grafen som textversionen är klickbara, så att man kan utforska lexikonets nätverksstruktur genom att klicka sig från lexem till lexem. Figur 2 visar den grafiska presentationen av lexemet *lexikon*[1] i gränssnittsprototypen.

FIGUR 2: *Grafiskt SAL-gränssnitt*



3. Användningar av SAL

Vad kan man tänka sig för användningar av SAL när den så småningom blir tillgänglig via Språkbanken? Bara ens egen fantasi sätter ytterst gränserna för vad man kan göra, naturligtvis, och här ska jag bara försöka ge ett litet axplock av möjliga användningar.

Först och främst är SAL intressant som lexikon, eftersom det representerar ett tidigare oprövat sätt att organisera språkets ordförråd semantiskt (åtminstone i den skala som det sker i SAL). I och med att det blir vidare tillgängligt, kan också lexikologer jämföra SAL med andra typer av semantiskt organiserade lexikon, t.ex. mer traditionella tesaurusar som Bring (1930) eller de lexikalisk-semantiska relationerna i NEO. Även mer ”vanliga” lexikonanvändare skulle kunna experi-

mentera med SAL som komplement till andra typer av semantiskt organiserade lexikon för användningar som ex.vis författande eller korsordslösning, två användningar som redan Lönngren (1992: VI) föreslår i förordet till den andra utgåvan av SAL, men som hittills inte har förverkligats i någon större utsträckning, p.g.a. lexikonets begränsade spridning.

Ur min egen synvinkel är emellertid de språkteknologiska tillämpningarna intressantast. En del av dessa – speciellt de som mer har karaktären av grundforskning – torde dock vara intressanta även för ”rena” lexikologer. Här erbjuder språkteknologin helt enkelt nya forskningsredskap för att belysa lexikologiskt intressanta frågeställningar, t.ex. de nedan nämnda frågorna om det inbördes förhållandet mellan olika typer av lexikalisk-semantiska betydelserelationer, om dessa relationers koppling till textuella faktorer, liksom om förhållandet mellan central vokabulär och textfrekvens.

Rent allmänt kan man säga att ett stort allmänt och fritt tillgängligt maskinlexikon med olika sorters lingvistisk information saknas för svenska. Om man ser till att en sådan lexikonresurs kommer till stånd, kommer den mänskliga uppfinningsrikedomen att se till att den används på mångahanda och oförutsedda sätt, det framgår med all önskvärd tydlighet av fallet (engelska) WordNet. Målet med det arbete som beskrivs här är som sagt just bland annat detta, att skapa en fri (för forskning och utbildning) omfattande språkteknologisk lexikonresurs med formaliserad lexikalisk-semantisk, böjningsmorfologisk och så småningom också syntaktisk information.

En sådan resurs skulle kunna användas som lexikonkomponent i språkteknologiska tillämpningar för olika ändamål, t.ex. i skrivstöd för att föreslå alternativa uttryck för skribenten att använda eller för att ”navigera” i stora textmaterial på semantisk basis genom att låta SAL generera semantiskt relaterade alternativa sökord som sen används i ett vanligt korpussöksystem.

Vidare finns en rad specifika språkteknologiska forskningsproblem där man skulle kunna ha god nytta av en stor lexikalisk-semantisk resurs för svenska, t.ex. problemet med att automatiskt lösa upp koreferens i text, som ofta sker med semantiskt relaterade nominalfraser (av typen *en bil ... fordonet*, etc.).

Det skulle också vara intressant att testa med automatiska metoder hur SAL:s olika relationer förhåller sig till mer traditionella lexikalisk-semantiska relationer eller till de associationer som man kan få fram ur stora korpusar med olika typer av maskininlärningsmetoder och som man tror sig veta väl fångar semantiska associationer mellan lexem (se t.ex. Rapp 2004).

En grundprincip i SAL:s uppbyggnad är ju att de primitiva semantiska relationerna går från mindre centrala till mer centrala lexem (dvs. den obligatoriska mödrarelationen). I många sammanhang brukar man approximera centralitet med frekvens. Därför skulle man också vilja jämföra SAL-lexemens avstånd från PRIM med lexemfrekvensdata från en balanserad svensk korpus, helst en med både tal- och skriftspråksdata. Nu finns det ingen sådan korpus, men vi har tillgång till frekvensdata från SUC (*Stockholm Umeå Corpus*; Ejerhed & Källgren 1997), som är en balanserad korpus av svenskt publicerat skriftspråk om en miljon ord. Ur SUC kan man emellertid bara få fram grundformsfrekvenser som kan differentieras med avseende på ordklass, så man får en approximation av lemmafrekvenser, men alltså inte lexemfrekvenser, eftersom samma lemma kan motsvara mer än ett lexem. Ungefär två tredjedelar av SAL-lexemen motsvarar dock bara ett lemma, så man skulle kunna använda dem för att testa om man finner någon sådan korrelation mellan SAL-lexemnivå och SUC-frekvens.

4. Om framtiden

Arbetet med den elektroniska SAL-versionen pågår som bäst, och vi räknar med att kunna lägga ut en första allmänt tillgänglig version av lexikongränssnittet i Språkbanken någon gång under hösten 2005. Den ska ge tillgång till samma information som pappers-SAL, men både i textformat och grafiskt format. Dessutom får man också lemma, ordklass och nivåangivelse för varje lexem.

Lagringsformatet och gränssnittet har båda utformats för att vara utbyggbara och återanvändbara. I kommande versioner av gränssnittet planerar vi att lägga till åtminstone följande information och funktioner:

- formaliserad morfologisk information för användning i automatiska analys- och genereringstillämpningar, t.ex. lemmatisering av korpusar
- lexikalisk-semantiska relationer från andra lexikon, ex.-vis NEO-databasen
- nya sökfunktioner, t.ex. sökning efter kortaste vägen mellan två lexem
- fler lexikon tillgängliga via samma gränssnitt och också länkade till varandra
- lemmafrekvenser hämtade ur SUC och andra korpusar
- koppling mellan lexikongränssnittet och korpussöksystem för sökning efter semantiskt besläktade ord i korpusar

- möjligheter för användare att lägga till egen information i databasen
- ett API för programmatisk uppslagning i SAL och andra lexikon över internet – en ”lexikonserver”

5. Litteraturförteckning

Bontcheva, Kalina, Atanas Kiryakov, Hamish Cunningham, Borislav Popov & Marin Dimitrov 2003. Semantic Web enabled, open source language technology. *Language technology and the Semantic Web – Workshop on NLP and XML (NLPXML-2003)*, held in conjunction with EACL 2003. Budapest: ACL.

Bring, S. C. 1930. *Svenskt ordförråd ordnat i begreppsklasser*. Stockholm: Hugo Gebers Förlag.

Broekstra, Jeen 2005 *Storage, querying and inferencing for Semantic Web languages*. SIKS Dissertation Series 2005–09. Vrije Universiteit, Amsterdam.

Ejerhed, Eva & Gunnel Källgren 1997. Stockholm Umeå Corpus Version 1.0, SUC 1.0. Department of Linguistics, Umeå University.

Ide, Nancy, Alessandro Lenci & Nicoletta Calzolari 2003. RDF instantiation of ISLE/MILE lexical entries. *Proceedings of the ACL'03 workshop on linguistic annotation: Getting the model right*. Sapporo: ACL. S. 30–37.

Forsberg, Markus & Aarne Ranta 2004. Functional morphology. *ICFP'04, Proceedings of the ninth ACM SIGPLAN international conference of functional programming*. Snowbird, Utah.

Lönngren, Lennart 1988. Svenskt associationslexikon. Rapport UCDL-R-88-2. Centrum för datorlingvistik. Uppsala universitet.

Lönngren, Lennart 1989. Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi. Rapport UCDL-R-89-1. Centrum för datorlingvistik. Uppsala universitet.

Lönngren, Lennart 1992. Svenskt associationslexikon. Del I–IV. Rapport från Institutionen för lingvistik. Uppsala universitet.

Lönngren, Lennart 1998. A Swedish associative thesaurus. *Euralex '98 Proceedings, Vol. 2*. S. 467–474. (Tillgänglig på WWW: <http://www.hum.uit.no/a/lonngren/Assoc.pdf>; läst 2005-06-30.)

Nationalencyklopedins ordbok 1995–96. Höganäs: Bra Böcker.

Rapp, Reinhard 2004. A freely available automatically generated thesaurus of related words. *Proceedings of LREC 2004*. Lissabon: ELRA. S. 395–398.

Svensk ordbok 1986. Stockholm: Esselte Studium.

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