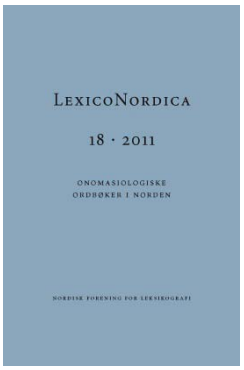
LexicoNordica

Titel: Swesaurus – ett svenskt ordnät med fria tyglar

Forfatter: Lars Borin & Markus Forsberg

Kilde: LexicoNordica 18, 2011, s. 17-40

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



Swesaurus – ett svenskt ordnät med fria tyglar

Lars Borin & Markus Forsberg

Swesaurus is a free Swedish wordnet currently under construction in Språkbanken, University of Gothenburg. Swesaurus is constructed by reusing information about lexical-semantic relations in a number of pre-existing freely available lexical resources: SALDO, SDB, Synlex and Swedish Wiktionary. In this article we describe how this existing lexical-semantic information is reused and enriched by utilizing the connections among the lexical resources. We focus in particular on our experiments with turning the graded synonymy relations of Synlex into *fuzzy synsets* in Swesaurus. The introduction of fuzziness into a wordnet raises many intricate methodological and theoretical questions, which are discussed and illustrated in the article.

1. Introduktion

Nedan beskriver vi ett pågående arbete för att skapa ett fritt svenskt ordnät. Detta görs till stor del genom återanvändning av ett antal befintliga fria lexikonresurser. På grund av strukturen hos dessa resurser får det nya ordnätet några i sammanhanget ovanliga och intressanta karakteristika, bland annat associativa relationer mellan ord och variabel, kvantifierad synonymi, av vilka den sistnämnda står i fokus i denna artikel.

1.1. Ordnät

De s.k. *ordnäten* har på relativt kort tid blivit centrala lexikonresurser för språkteknologitillämpningar. Det första ordnätet var det

engelska *Princeton WordNet* (PWN), som började utvecklas i mitten av 1980-talet av en forskargrupp vid Princeton-universitetet i USA under ledning av George A. Miller (Fellbaum 1998; PWN 2011). Idag finns ordnät för ett antal språk (se GWN 2011), och många av dem är fritt tillgängliga under samma typ av licens som det ursprungliga PWN, som tillåter alla slags användningar liksom vidareutveckling och återpublicering utan några restriktioner.

1.2. Ordnätens idé och organisation

Ordnäten är onomasiologiska lexikonresurser. Deras uppbyggnad motiverades ursprungligen inte av lexikografiska eller teoretisk-lingvistiska hänsyn, utan av psykolingvistiska och kognitionsvetenskapliga idéer om det mentala ordförrådets organisation. Grundenheterna i ett ordnät är något som man skulle kunna kalla *lexikaliserade begrepp*, men som man normalt utan förlust av precision kan tänka på som ordbetydelser – ett Saussureanskt språkligt tecken, alltså en kombination av språkligt uttryck (ett lexikonord) och innehåll. Ordbetydelserna organiseras i synonymmängder (synsets) enligt en speciell definition av synonymi (se avsnitt 2). Synonymmängder och lexikonord kopplas samman med traditionella – i litteraturen om ordnät ofta kallade ”klassiska” – lexikalisk-semantiska relationer, av vilka den viktigaste, utöver synonymi, är hyponymi och därmed automatiskt dess omvändning hyperonymi. En lexikonresurs måste minimalt ha synonymmängder hierarkiskt organiserade av hyponymirelationer för att räknas som ett ordnät (GWN 2011).

PWN innehåller enbart de fyra öppna ordklasserna substantiv, verb, adjektiv och adverb. På grund av vad som verkar vara en specifik anglosaxisk lexikografisk tradition (Apresjan 2002) finns dock även räkneord med i PWN, klassificerade som substantiv och adjektiv.

1.3. Mot ett fritt svenskt ordnät: *Swesaurus*

För svenska finns för närvarande inte något ordnät som är tillgängligt med en liknande typ av licens som PWN, vilket hämmar utvecklingen av en rad språkverktyg för svenska som är självklart tillgängliga för engelska.[1]

Swesaurus syftar till att fylla den luckan för svenskans vidkommande. Det är ett svenskt ordnät som för närvarande är under utveckling i Språkbanken (SB 2011) vid Göteborgs universitet och som görs tillgängligt under en typ av öppen källkodslicens. Swesaurus är delvis ett resultat av kombinationen av redan existerande, fria resurser. De fyra centrala befintliga resurserna är SALDO, SDB, Synlex och Wiktionary. I avsnitt 3 beskriver vi dessa resurser och hur informationen i dem används för att bygga upp Swesaurus. Först behöver vi dock diskutera den i det här sammanhanget centrala termen *synonymi* lite närmare, vilket vi gör i nästa avsnitt.

2. Synonymi

Synonymi är den centrala lexikalisk-semantiska relationen i ett ordnät. Synonymi definieras så här i PWN: Om två ord är utbytbara i minst en mening utan att meningens betydelse – i termer av sanningsvillkor – förändras, så uttrycker de två ordbetydelserna samma begrepp och bildar därmed en synonymmängd. Det grundläggande arbetsmomentet när ett ordnät ska byggas är alltså att bestämma vilka ordbetydelser som är synonyma i den här bemärkelsen.

1 I början av 2000-talet bedrevs i Sverige ett projekt finansierat med offentliga medel för att bygga en svensk motsvarighet till Princeton WordNet. Projektet resulterade i ett svenskt ordnätsfragment med i storleksordningen 25.000 synonymmängder. Detta ordnät verkar inte ha utvecklats vidare sedan mitten av 2000-talet och det är explicit inte tillgängligt under villkor som skulle göra det möjligt att bygga vidare på det i arbetet med Swesaurus.

Det här är ett rätt speciellt sätt att förstå synonymi och en intressant fråga som uppkommer i det sammanhanget är huruvida man i olika ordnätsprojekt förstår synonymi på samma sätt som PWN:s skapare. Svaret är faktiskt nej, åtminstone när det gäller det danska ordnätet DanNet (Pedersen et al. 2009) och det ena av de två polska ordnäten (Piasecki et al. 2009). En förklaring till det kan vara att de två nämnda projekten – i motsats till PWN och i motsats till ordnät som har skapats genom översättning av PWN, som det finska ordnätet (Lindén & Carlson 2010) – har utgått ifrån starka lexikografiska traditioner där man redan av hävd har en annan idé än i PWN-projektet om vad som inryms i begreppet synonymi. PWN-definitionen av synonymi – eller snarare dess konsekvenser – att synonymi används för att definiera den grundläggande lexikonenheten, så att man får lika många lexikonenheter som man kan hitta (på) utbyteskontexter – känns i själva verket lite främmande mot en sådan bakgrund. Dessutom kan man notera en vacklan även vad beträffar PWN. Även om man i litteraturen alltid framhåller synonymmängden (synset) som den grundläggande enheten i PWN och andra ordnät, kan detta inte vara hela sanningen: mer än hälften av alla synonymmängder i PWN 3.0 (54 %) har bara en medlem – som därmed rimligen inte uppvisar utbytbarhet mot något annat ord i en enda kontext – varför man måste tänka sig att det finns ytterligare sätt att avgöra vilka ordbetydelser man har att räkna med i ett ordnät. Inget annat ordnät torde ha lika stor betydelsesplittring som PWN, där (grundformen) *break* i PWN 3.0 innehar rekordet med 75 betydelser, fördelade på 59 verb och 16 substantiv.[2] En sådan stor mängd

2 Vår jämförelsegrund är dock uttryckligen andra ordnät och andra lexikonresurser för språkteknologi. Jämfört med konventionella engelska lexikon är PWN inte anmärkningsvärt. Trots sina psykolingvistiska anspråk följer PWN här som i en del andra avseenden helt enkelt en engelskspråkig lexikografisk tradition: stora engelska konventionella lexikon anger ännu fler betydelser för break (107 betydelser i DICTIONARY.COM 2011, varav 71 verb och 36 substantiv).

betydelser är konstaterat svår att differentiera maskinellt, vilket har konsekvenser för automatisk textanalys som använder sig av PWN.

Synonymordlistor har gjorts lika länge som man överhuvudtaget har sammanställt ordlistor och ordböcker; redan sumerer och akkadier lämnade efter sig synonymlistor (Civil 1990) och latinsynonymiken var en på sin tid blomstrande lexikongenre som gick tillbaka till det första århundradet av vår tideräkning och som fortfarande var levande under det sena 1800-talet, en tidsrymd om nästan två årtusenden (Svensén 2010). Det är alltså inte ägnat att förvåna att synonymi med tiden har kommit att förstås på många olika sätt. Ytterligheterna illustreras väl med citat ur två olika latinsynonymiker från artonhundratalets mitt:

> Still it ought to be remembered, that there are actually words which differ simply as to form or sound, and by which the scholar is strangely misled, if he starts with the axiom, that there are no two words meaning exactly the same thing, in the same language, – an error, it seems, which may be perceived in far the greater number of works on synonymes. That there are equivalent words may be seen at once, if we remember, that some Latin words end both in is and us, without a shadow of difference in meaning. (Lieber 1841:vii)

> If, therefore, it be admitted, that words identical in meaning do not exist, and that it is morally impossible, if I may use the expression, that they should exist, the only questions are, whether, in such cases, it is worth while to search out their differences, and whether it is possible to find them out. Science will answer the first question, without hesitation, in the affirmative; and with respect to the second, there can at least be no presumption in making the attempt. (Döderlein 1863:xii)

Å ena sidan är det självklart att det finns ord med exakt samma betydelse (Lieber), å den andra är det ”moraliskt omöjligt” att sådana ord skulle existera (Döderlein). I det här avseendet har ingenting ändrats i sak sedan 1800-talet. Även om den lingvistiska argumentationen har blivit mycket mer sofistikerad, finns det lika många bud nu som då om hur synonymi ska förstås och extremståndpunkterna är fortfarande att synonymi är antingen icke-existerande eller ofrånkomlig (se Murphy 2003, kap. 4).

I praktiken verkar både lexikografer och många ordnätsmakare – som i de ovan nämnda danska och polska ordnätsprojekten – använda en mer traditionell definition av synonymi än den som används i PWN, nämligen en där lexikonets ord(betydelser) fastställs först, på andra, oberoende grunder, och synonymi sen blir en relation mellan sådana lexikonord, som oftare i själva verket motsvarar partiellt snarare än totalt sammanfall i betydelse. Mot den bakgrunden kunde det vara intressant att försöka kvantifiera graden av överensstämmelse i betydelse mellan synonymer i denna bemärkelse. Det är möjligt att man skulle kunna företa en sådan kvantifiering med PWN, genom att räkna antalet synset som två ord uppträder i samtidigt (om det finns sådana synset). Såvitt vi känner till har detta dock inte gjorts.

En sådan kvantifiering öppnar för möjligheten att skapa ordnät där synonymmängderna kan bli större eller mindre beroende på hur strikt man förstår synonymi. Därmed skulle man kunna komma åt något som ofta har påpekats som ett praktiskt problem med PWN, nämligen att det har så fin betydelseindelning att de metoder man för närvarande förfogar över för automatisk textanalys inte klarar av att tilldela PWN-betydelser till ordförekomster i texter. Ofta efterfrågas i sådana sammanhang ett principiellt välgrundat och pålitligt sätt att slå ihop PWN-betydelser (och därmed få färre synonymmängder). PWN:s struktur erbjuder ingen hjälp därvidlag, eftersom de betydelser som hör till samma grundord inte är hierarkiskt ordnade i huvudbetydelser och underbetydelser. Se t.ex. Palmer et al. (2006) och referenser där.

3. Från SALDO, SDB, Synlex och Wiktionary till Swesaurus

I detta avsnitt presenterar vi de fyra resurser som ligger till grund för Swesaurus, tillsammans med en översiktlig beskrivning av arbetet med att inkludera de lexikalisk-semantiska relationerna från dessa resurser i Swesaurus.

Vi nämnde i avsnitt 1 ovan att dessa resurser är fria, med vilket vi helt enkelt menar att de har användningsvillkor (licenser) som tillåter vem som helst att använda resurserna till vilket ändamål som helst (inklusive kommersiell användning), samt även modifiera eller vidareutveckla och återpublicera dem, med det enda villkoret att de modifierade och återpublicerade resurserna inte får beläggas med restriktivare användningsvillkor än de ursprungliga resurserna. Detta är vad som i mjukvaruvärlden kallas *öppen källkod* (’open source’) och vad man ibland för andra typer av resurser än datorprogram kallar *öppet innehåll* (’open content’).[3] Om man som vi vill att slutprodukten ska vara fritt tillgänglig på detta sätt, måste även alla ingående delresurser vara det. Detta begränsar naturligtvis vilka andra lexikonresurser vi kan använda i arbetet med Swesaurus, men det är essentiellt för vårt forskningsområde att vi kan bygga den typ av öppen språkteknologiinfrastruktur som PWN och andra fria resurser är goda exempel på. Medaljens baksida kan tyckas vara att vi därmed måste avstå från att använda högvärdiga lexikaliska databaser som av kommersiella och andra skäl inte är fritt tillgängliga i den mening som avses här. Vi måste naturligtvis respektera att många vill ha kontroll över sina lexikondata av olika skäl. Samtidigt behöver svensk språkteknologi fria resurser för att växa, vilket betyder att det egentligen i praktiken inte finns något val: ur vårt forskningsområdes synvinkel existerar helt enkelt inte dessa resurser.

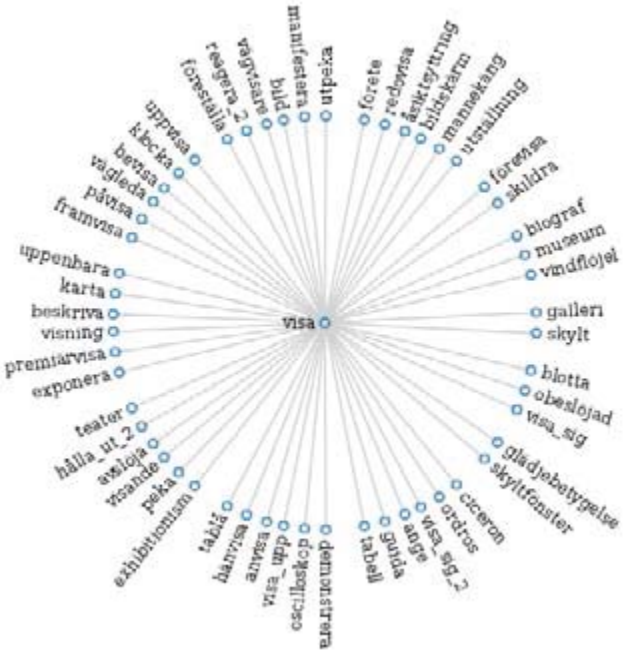
3 Det finns en hel rad sådana licenser; de som vi använder för våra lexikonresurser heter GNU LGPL och Creative Commons BY-SA.

3.1. SALDO

SALDO är en fullskalig (>100.000 ordbetydelser) lexikalisk-semantisk – onomasiologisk – resurs för svenska. SALDO uppdateras och utvidgas kontinuerligt. Arbetet med SALDO sker väsentligen i öppenhet: utvecklingsversionen publiceras dagligen dels som nedladdningsbar resurs, dels via en uppsättning sökgränssnitt (SALDO 2011). De nedladdningsbara utvecklingsversionerna görs tillgängliga via ett versionshanteringssystem, ett system som möjliggör att varje enskild version går att återskapa. Detta är en viktig egenskap i en resurs även avsedd för användning inom forskning, där reproducerbarhet är ett centralt metodologiskt krav.

Grundenheterna – (lexikon)ingångarna – i SALDO är ordbetydelser som ordnas genom en grundläggande semantisk relation som vi kallar *association*. I likhet med PWN innehåller SALDO en stor mängd flerordsenheter och egennamn, men till skillnad från PWN beskriver SALDO även ord ur de slutna ordklasserna (pronomen, prepositioner, subjunktioner, etc.). Varje ordbetydelse beskrivs genom en association i form av en *primär deskriptor* eller 'moder'. Den primära deskriptorn ska vara den semantiskt mest närliggande ordbetydelse som också är centralare än den ordbetydelse som ska beskrivas. Det senare kriteriet tillsammans med kravet att alla deskriptorer i sin tur ska ha en beskrivning i SALDO (med ett undantag; se nedan) ger SALDO en hierarkisk struktur. Kriterierna är inte formellt definierade, utan den som arbetar med att definiera nya ingångar i SALDO får använda sin språkliga intuition och sitt omdöme. Vad gäller centralitet finns dock några typiska kännetecken, såsom att en mer central ordbetydelse har *högre korpusfrekvens*, är *morfologiskt enklare*, är *stilistiskt neutralare*, är *mindre ämnesbunden* och *uppträder tidigare i barns språkutveckling* än en mindre central betydelse. Några exempel på primära deskriptorer är *pojke* för *kille*, *sannolik* för *sannolikhet*, och *leva* för *liv*.

När det gäller att avgöra vilken den mest närliggande ordbetydelsen är, visar praktiken att den ständiga interaktionen med lexikonmaterialet ger en bra vägledning. Ordbetydelser med samma primära deskriptor bör när man ser dem samlade bilda en tydlig 'semantisk familj', en gestalt, och det är relativt lätt att se om något eller några ord sticker ut ur mängden (se figur 1). SALDO kan ses som resultatet av många lokala beslut, som dessutom ständigt revideras.



Figur 1: Ordbetydelser i SALDO med *visa* som primär deskriptor.

Det finns ett litet antal – 43 stycken – ordbetydelser som inte har någon primär deskriptor i SALDO. Dessa förses med en artificiell 'rotbetydelse' eller 'nollbetydelse' kallad *PRIM* för att SALDO formellt ska bilda en enda trädstruktur (istället för 43 trädstrukturer). PRIM har ingen beskrivning i SALDO.

Över hälften av alla ordbetydelser i SALDO har enbart den primära deskriptorn, men för att differentiera betydelsebeskrivningen kan man ange en eller flera *sekundära deskriptorer* eller 'fäder'. Exempelvis har både *menageri* och *anekdotflora* den primära deskriptorn *samling*, men de skiljs åt i den sekundära deskriptorn: *djur* för *menageri*, *anekdot* för *anekdotflora*. Av de sekundära deskriptorerna krävs inget annat än just att de hjälper till att skilja åt ordbetydelser som delar primär deskriptor. Därför är de sekundära deskriptorerna ofta semantiskt mindre centrala än den betydelse som ska beskrivas. Ungefär en femtedel (strax över 9.000) av de sekundära deskriptorerna är mindre centrala – har en primär deskriptor längre bort från PRIM – än den betydelse de beskriver.

En primär deskriptor är ofta en hyperonym eller synonym (eller i sällsynta fall en hyponym) till den beskrivna ordbetydelsen. Därför kan SALDO användas vid uppbyggnaden av Swesaurus. Det är enkelt att ta ut alla par i SALDO där lexikoningången och dess primära deskriptor har samma ordklass. Specialfallet sammansättningar, där efterledet i sammansättningen typiskt är dess primära deskriptor, kan särbehandlas genom att mekaniskt kontrollera att lexikoningångens grundform slutar med den primära deskriptorns grundform. På detta sätt samlar man synonym/hyponymkandidatlistor som snabbt kan gås igenom för hand. Detta är ett arbete som är påbörjat.

Swesaurus kommer som en slags bonus även att innehålla de associativa relationerna från SALDO, vilka kan inkluderas utan något extra arbete eftersom en enda uppsättning formella ordbetydelseidentifierare – nämligen SALDOs – används i alla våra lexikonresurser (Borin 2010). Det betyder att Swesaurus redan från början får en typ av relationer gratis som man nu, intressant nog, lägger ner stort arbete på att lägga till i Princeton WordNet och som man i det sammanhanget kallar för ”framkallande” (eller ”frammanande”: ”evocation”; Boyd-Graber et al. 2006; se även Morris & Hirst 2004).

SALDO följer i stort sin föregångare SAL när det gäller den semantiska organisationen. Den som vill ha en mer ingående beskrivning av SALDO hänvisar vi till Borin (2005) för den semantiska organisationen och Borin et al. (2008) för övriga aspekter av SALDO.

3.2. SDB

SDB (semantisk databas; Järborg 2001) är en lexikalisk databas med 68.000 ordbetydelser, av vilka ett antal är mycket detaljerat beskrivna med avseende på sin semantiska valens, som anges med användning av ett fyrtiotal semantiska roller. Många av ordbetydelserna är också länkade till ordförekomster i en svensk standardkorpus. Delar av informationen i SDB ska användas vid uppbyggnaden av ett svenskt frasnät (Borin et al. 2010).

För vårt arbete med Swesaurus är dock det intressantaste med SDB att resursen tillhandahåller information om en rad klassiska lexikalisk-semantiska relationer mellan ett stort antal av sina lexikoningångar: synonymi, hyponymi, kohyponymi, antonymi och en 'allmän semantisk relation' ("se"). Den sistnämnda anger helt enkelt något slags semantiskt samband, utan att försöka specificera det närmare, inte olikt SALDOs associationer, men inte nödvändigtvis med samma krav på att peka på en centralare betydelse. Ett exempel är golftermen *albatross* i betydelsen "tre slag under par" som står i "se"-relation till en annan golfterm, *par*.

En relation kan vara explicit angiven eller logiskt härledbar. Exempelvis betraktas ofta synonymi som en *diskret* – till och med *binär – transitiv relation*. Att relationen är binär betyder att den antingen råder eller inte; man räknar inte med några grader av synonymi.

Transitivitet betyder i den tekniska bemärkelse som vi använder den här följande. Om vi vet att en viss transitiv relation råder mellan objekten A och B och att samma relation råder mellan ob-

jekten B och C, så kan vi därmed sluta oss till att relationen råder även mellan A och C. Transitivitet kan alltså sägas vara den formella versionen av en utsaga som ”Min väns vän är min vän”. En prototypisk transitiv relation i den fysiska världen är ”större än”. Om vi får veta att en elefant är större än en kamel och att en kamel är större än en koala, så vet vi också automatiskt att en elefant är större än en koala, utan att detta behöver sägas explicit.

Att synonymi är transitiv innebär alltså att om vi vet att A är synonymt med B och B med C, så följer med automatik att A är synonymt med C, eller med andra ord, synonymparet A–C kan härledas från de explicit angivna synonymparen A–B och B–C. Den mängd man får om man gör detta uttömmande för en transitiv relation kallas *det transitiva höljet*, ett begrepp som kommer att vara centralt när vi nedan diskuterar användningen av graderade synonympar ur Synlex (se avsnitt 4).

Det transitiva höljet är helt enkelt den mängd av objekt som man får om man explicit räknar ut alla transitivitetsrelationer givet något eller några av objekten. Om vi igen använder exemplet ”Min väns vän är min vän”, så får jag det transitiva höljet genom att först räkna upp mina vänner, för var och en av dem hennes eller hans vänner, för var och en av dem hans eller hennes vänner, etc., tills jag har uttömt alla möjligheter. Mängden av individer som jag har fått fram på det sättet är det transitiva höljet med avseende på relationen ”vän med” och utgående från mig själv.

I vårt fall kan man tänka på det transitiva höljet som en ’kedja’ av ordbetydelser där varje länk i kedjan ges av ett synonympar, eller med ett konkret exempel, det transitiva höljet för synonymparen *barn–parvel*, *parvel–pys* och *pys–knatte* är synonymmängden {*barn*, *parvel*, *pys*, *knatte*}.

Synonymi är även en *symmetrisk* relation: om A är synonymt med C, så följer även att C är synonymt med A. Hyponymi och hyperonymi är varandras *inverser*: om A är en hyponym till B, så är B en hyperonym till A.

Med kännedom om hur de lexikalisk-semantiska relationerna förhåller sig logiskt till varandra, kan man använda dem och deras kombinationer för att 'fylla på' ofullständig information hämtad ur en lexikonresurs som SDB, där inte alla logiskt möjliga synonym- och hyponym/hyperonympar anges explicit, utan bara ett litet fåtal av dem.

Den allmänna strategin för att använda den lexikalisk-semantiska informationen i SDB blir då följande: (1) extrahera relevanta ordbetydelsepar i SDB; (2) beräkna transitiva höljet med avseende på synonymi, samt beräkna symmetri och invers för att därmed utöka till alla möjliga par; (3) undersöka resultatet manuellt för att hitta iögonenfallande fel. Detta arbete är påbörjat.

3.3. Synlex

Synlex (Kann & Rosell 2006; SYNLEX 2011) är en graderad svensk synonymlista skapad av Viggo Kann och hans kollegor vid KTH i Stockholm. De har utvecklat ett system där användare av ett svenskt-engelskt nätlexikon ombeds att bedöma graden av synonymi för ordpar som genererats automatiskt ur tvåspråkiga lexikon och ur korpusar, på en skala från 0 (ej synonyma) till 5 (fullt synonyma). Paren är ordformer utan ordklassangivelse, huvudsakligen grundformer, men även böjda former och flerordsuttryck förekommer. Några exempel på ordpar tagna direkt från Synlex system är *stark–högaktiv*, *fullständig–rätt*, *ram–gränsområde* och *design–mönster*.

En fritt nedladdningsbar delmängd av Synlex publiceras med oregelbundna mellanrum på projektets hemsida (SYNLEX 2011). Datamängden innehåller ordpar som har fått ett minimum av tre bedömningar och som har en genomsnittlig synonymigrad om minst 3.[4] Versionen daterad 2010-09-19 innehöll c:a 19.000 grade-

4 Synlexgraderingen 3–5 multipliceras med 20 i Swesaurus, vilket ger ett intervall 60–100 som man med fördel kan tänka på som en procentangivelse, där 100 representerar full synonymi.

rade synonympar (det dubbla om man tar hänsyn till synonymirelationens symmetri).

Hittills har vårt arbete med Swesaurus fokuserat mest på just Synlex som informationskälla. Detta arbete beskrivs utförligare i avsnitt 4.

3.4. Wiktionary

Wiktionary är en ansats liknade Wikipedia men för kollaborativt författande av nätlexikon snarare än av nätuppslagsverk. Svenska Wiktionary (WIKTIONARY 2011) är en nedladdningsbar fri resurs som – i tillägg till mycket annat – innehåller en hel del lexikalisk-semantiska relationer. Arbetet med att extrahera sådana relationer ur Wiktionary försvåras dock av att datamängden är enbart delvis formellt strukturerad. Det kommer nämligen an på författaren av en lexikoningång att använda det givna wikiformatet för att koda de olika informationskategorierna i ingången på det sätt som Wiktionarys upphovsmän har avsett, men det görs ingen kontroll av att så faktiskt sker. Eftersom resultatet av en felaktig kodning kan se rätt ut för det mänskliga ögat, finns det i praktiken en rad formella fel i Wiktionary som komplicerar den automatiska informationsutvinningen.

Vi har experimenterat med att extrahera synonymirelationer mellan ord, vilket hittills resulterat i 4.012 synonympar, varav 1.514 är ordpar där vardera medlemmen i paret motsvarar endast en betydelse i SALDO. Därmed behövs ingen manuell disambiguering, utan dessa kan införlivas i Swesaurus direkt. En del av synonymparen i Wiktionary är felaktiga, vilket kommer sig av att vissa ordbeskrivningar även innehåller information om andra språk och relationer inom dessa. Detta resulterar i ett fåtal fall där exempelvis ett svenskt ord länkas till ett polskt. I praktiken innebär detta inget problem, eftersom länkningen till SALDO även innebär att ord som inte förekommer i SALDO filtreras bort.

Synonymirelationerna i Wiktionary är generellt av högre kvalitet än i Synlex, vilket är att förvänta sig eftersom författaren av en lexikoningång i Wiktionary gör ett aktivt val när hon anger synonymer till det ord som beskrivs, medan Synlex bygger på bedömningar av automatiskt genererade ordpar, vilket resulterar i att ord som vanligtvis inte anses vara synonyma ändå kan komma att få en gradering större än noll. Ta ord som *förlovning* och *förpliktelse*, vilka vanligen inte anses synonyma, men när man får dem presenterade tillsammans med en uppmaning att kvantifiera synonymigraden mellan dem, lockas man nog till att ange ett visst mått av synonymi. Medaljens baksida är att det finns betydligt färre synonympar i Wiktionary, vilket också är det förväntade.

Det svenska Wiktionary har som tidigare nämnts ordbeskrivningar som innehåller lexikal information om andra språk, i huvudsak engelska. Det finns översättningslänkar mellan svenska och engelska som kan utnyttjas för att skapa ett svenskt-engelskt lexikon. Ett arbete som är knappt påbörjat är att undersöka om den svensk-engelska informationen kan användas för att skapa en första approximation till länkning av Swesaurus till Princeton WordNet.

4. Graderad synonymi och luddiga synonymmängder

Som vi nämnde i avsnitt 2 är det som kallas synonymi normalt närsynonymi, så ett försök att gradera synonymi som i Synlex tycks intuitivt korrekt. Det finns i princip ett otal tänkbara sätt att beräkna denna synonymigrad beroende på vår uppfattning om vad synonymi är och vad vi anser vara en ordbetydelse. Men varför inte försöka använda synonymigraderna i Synlex, eftersom de är fritt tillgängliga?

Frågan är hur man förenar Synlex graderade synonymi med

det synonymibegrepp som konstituerar ett ordnät som PWN, som ju består av synonymmängder och lexikalisk-semantiska relationer mellan dessa mängder. Som vi redan har nämnt så ses synonymi i många sammanhang – inklusive PWN – som en binär, transitiv relation. Därmed ges en synonymmängd i normala fall av det *transitiva höljet* (se avsnitt 3.2).

Frågan kompliceras dock av graderingen, alltså att vi oftast har att göra med närsynonymi snarare än total synonymi, eftersom det inte är självklart att närsynonymi är transitiv. Om A är en närsynonym till B med graden *x* och B till C med graden *y*, följer det då med nödvändighet att A är närsynonym till C, och i så fall med vilken grad? Ett exempel taget från Synlex är ordparen *precis–absolut* med graden 60, och *absolut–fullkomlig* med graden 62, där frågan vi ställer oss är om det därmed är givet att *precis–fullkomlig* är närsynonymer, och om så är fallet, med vilken grad? Införandet av graderad synonymi i ett ordnät får också konsekvenser för andra relationer, exempelvis för hyperonymi–hyponymi och holonymi–meronymi, eftersom dessa relationer i PWN:s lexikalisk-semantiska modell råder mellan synonymmängder och inte mellan ord. En annan fråga man därmed måste ställa sig är hur graderad synonymi påverkar dessa andra relationer. Exempelvis, om A är synonym med B med graden x, och C är en hyperonym till A, är det då nödvändigtvis så att även C är en hyperonym till B? Det enda sättet att tackla dessa frågor är att empiriskt testa olika antaganden och granska konsekvenserna.

4.1. Synonymmängder som kedjor av synonympar

Vår första ansats var att experimentera med transitiva höljet tillsammans med olika brytpunkter, där enbart synonympar med en gradering vid eller ovan brytpunkten tilläts bilda länkar i kedjan. Till exempel skulle en brytpunkt på 90 endast tillåta ordpar med en gradering på 90 eller mer. En intressant följd av att använda brytpunk-

ter är att det ger upphov till en mängd ordnät, ett ordnät per brytpunkt, eller vad man skulle kunna kalla ett parametriserat ordnät.

Resultatet blev å ena sidan en uppsättning av rimliga synonymmängder, men å den andra även en återstående ensam synonymmängd med flera tusen ordbetydelser. Denna onormalt stora synonymmängd uppstår på grund av: (1) vi har ett betydelsekontinuum; (2) Synlex innehåller ordpar som kanske inte ska räknas som synonymrelationer; (3) vissa ordbetydelser saknas i SALDO. Antagandet om entydighet mellan ordform och ordbetydelse i SALDO är naturligtvis giltigt endast under förutsättning att alla betydelser kan antas beskrivna i SALDO. Så är det förstås inte, utan i praktiken ser vi en växelverkan mellan de lexikonresurser som vi arbetar aktivt med (främst SALDO, Swesaurus och det svenska frasnätet), så att nya ordbetydelser för befintliga grundformer ständigt får läggas till i SALDO.

Vi lade sedan till kravet att en synonymmängd endast får bestå av ord i samma ordklass – något som ofta anses gälla för synonymi och som alltid gäller i PWN-modellen – vilket minskade den största synonymmängden, men antalet ordbetydelser i den kan fortfarande räknas i tusental.

4.2. Synonymmängder där alla känner alla

Vårt nästa experiment var en mer konservativ ansats: Vi krävde att en synonymmängd ska vara en *klick*. En synonymmängd är en klick när alla ingående ordbetydelser är (på förhand angivna som) synonyma med varandra, vilket sammanfaller med vad vi intuitivt menar med synonymmängd. För det transitiva höljet krävde vi ju bara en kedja av (på förhand givna) synonymirelationer från den första till den sista medlemmen i synonymmängden, och härledde sedan resterande relationer, t.ex. den mellan kedjans första och sista länk. Här kräver vi istället att alla synonymirelationer i mängden redan är explicit givna.

Om vi igen använder exemplet med ”Min väns vän är min vän”, så måste vi i detta fall lägga till ”men bara om hon själv säger både att hon är min väns vän och att hon är min vän”; vi får alltså nu bara samla ihop individer som alla säger sig vara vänner med alla andra i gruppen. Det inses lätt att den gruppen i normalfallet blir betydligt mindre än det transitiva höljet.

Beräkningen av klickar gav inga onormalt stora synonymmängder, men däremot ett par andra problem. Det första problemet är att ordbetydelser stundtals förekommer i fler än en synonymmängd, vilket strider åtminstone mot vad man i ordnätssammanhang brukar mena med ordbetydelse. Typiskt skiljer sig dessa synonymmängder åt endast med avseende på en eller ett par ordbetydelser – en indikation antingen på saknade synonymirelationer eller på saknade ordbetydelser i SALDO.

Naturligt nog uppstår då frågan om hur vi härleder saknade (när)synonympar, så att en ordbetydelse endast förekommer i en synonymmängdsklick. En del av det arbetet är att samtidigt identifiera saknade ordbetydelser.

En möjlig tanke är att härleda alla implicita par i det transitiva höljet som vi därefter går igenom manuellt, men det ger oss en ohanterlig mängd par. Den största synonymmängd som vi fick genom att använda det transitiva höljet innehöll 5.770 ordbetydelser. När varje ordbetydelse paras ihop med varje annan ordbetydelse i synonymmängden motsvarar det $5.770 \cdot 5.769 = 33.287.130$ synonympar. Antalet par kan förvisso halveras, eftersom om vi har A–B så behövs inte den omvända relationen B–A, men att ’bara’ behöva gå igenom knappt 17 miljoner synonymparskandidater manuellt hjälper oss inte mycket i sammanhanget.

En mer försiktig härledning av nya synonympar är att utgå från de klickar som har en eller flera ordbetydelser gemensamma och härleda de par som, om de existerade, skulle förena klickarna till en klick. Metoden kan upprepas över de sålunda sammanslagna klickarna, och därmed härleda nya par. Strategin ger upphov till

par av god kvalitet, speciellt om vi begränsar oss till det relativt vanliga fallet med klickar där alla medlemmar utom en är gemensam. Då leder begränsningen dock givetvis till att endast klickar av samma storlek kan förenas genom denna strategi.

5. Sammanfattning och utblick mot framtiden

Arbetet med Swesaurus bygger vidare på vad andra åstadkommit, vilket gör att vi hoppas kunna skapa en omfattande resurs av godtagbar kvalitet med en i sammanhanget relativt blygsam arbetsinsats.

Swesaurus blir således ett fritt svenskt ordnät med ett par i dessa sammanhang ovanliga egenskaper: Det kommer till en del att ha graderade synonymmängder, vilket i praktiken gör det till lika många ordnät som man väljer brytpunkter för graden av synonymi, och det kommer att ha SALDOs associativa relationer i tillägg till de klassiska lexikalisk-semantiska relationerna som vi förknippar med ett ordnät.

Det återstår ett antal teoretiska och metodologiska frågor att ta ställning till som har med graderingen att göra, som vi ännu inte rört vid. Särskilt kan här nämnas vilka effekter luddiga synonymmängder har på andra relationer, om några, och hur vi ska åstadkomma någon form av gradering för synonympar hämtade från någon annan resurs än Synlex. Empiriskt verkar åtminstone en tvågradig skala fungera för det praktiska arbetet med synonymparskandidater både ur SALDO (se avsnitt 3.1) och ur SDB (se avsnitt 3.2); det visar sig vara naturligt och relativt enkelt att klassificera dessa i *närsynonymer* (som tentativt åsätts 90 % synonymi) och (fulla) *synonymer* (och naturligtvis par som inte är någondera).

Vi har börjat utforska automatiska metoder för att göra kvalificerade gissningar om kvaliteten av existerande och härledda

par. En metod vi ska titta närmare på är om avståndet mellan två ordbetydelser i SALDOs hierarki kan användas för att göra detta. Detta är ett dubbelriktat arbete, eftersom om vi har ett stort avstånd, men synonymparet visar sig vara av god kvalitet, kan det i sin tur leda till en revidering av SALDO.

Vi planerar att släppa Swesaurus 1.0 i slutet av 2011, men redan nu är utvecklingsversionen sökbar och nedladdningsbar via Språkbankens lexikala system SBLEX (SBLEX 2011). Utvecklingsversionen uppdateras varje natt, vilket inbjuder till insyn och deltagande i vårt dagliga arbete. Om vårt arbete av någon anledning skulle upphöra, så kan dessutom därmed någon annan ta upp tyglarna och fortsätta precis från den punkt där vi släppt dem.

Litteratur

Apresjan, Yuri D. 2002: Principles of systematic lexicography. I: Marie-Hélène Corréard (red.): *Lexicography and natural language processing. A Festschrift in honour of B. T. S. Atkins*. Grenoble: Euralex, 91–104.

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

Borin, Lars 2010: Med Zipf mot framtiden – en integrerad lexikonresurs för svensk språkteknologi. I: *LexicoNordica* 17: 35–54.

Borin, Lars & Markus Forsberg 2009: 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. Odense: NEALT, 7–12.*

Borin, Lars, Markus Forsberg & Lennart Lönngren 2008: The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. I: Joakim Nivre, Mats Dahllöf and Beáta Megyesi (red.): *Resourceful language tech-*

nology. Festschrift in honor of Anna Sågvall Hein. Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7. Uppsala: Uppsala University, 21–32.

Borin, Lars, Dana Danélls & Markus Forsberg, Dimitrios Kokkinakis & Maria Toporowska Gronostaj 2010: The past meets the present in Swedish FrameNet++. I: *Proceedings of the 14th EURALEX International Congress*. Leeuwarden: EURALEX, 269–281.

Boyd-Graber, Jordan, Christiane Fellbaum & Daniel Osherson, Robert Shapire 2006: Adding dense, weighted connections to WordNet. I: *Proceedings of the Third International Wordnet Conference, GWC 2006*. Brno: Masaryk University, 29–35.

Civil, Miguel 1990: Sumerian and Akkadian lexicography. I: Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (utg.): *Wörterbücher: Ein internationales Handbuch zur Lexikographie. Zweiter Teilband / Dictionaries: An international encyclopedia of lexicography. Second volume / Dictionnaires: Encyclopédie internationale de lexicographie. Tome second*. Berlin: Walter de Gruyter, 1682–1686.

Döderlein, Ludwig 1863: The author's preface. I: *Döderlein's handbook of Latin synonymes*. Translated by Rev. H.A. Arnold, B.A., with an introduction by S.H. Taylor, LL.D. Andover: Warren F. Draper, ix–xvi.

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

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.

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

Lieber, Francis 1841: Preface of the translator. I: *Dictionary of Latin synonymes*, for the use of schools and private students, with a complete index. By Lewis [Ludwig] Ramshorn. From the German by Francis Lieber. Boston: Charles C. Little and James Brown, iii–viii.

Lindén, Krister & Lauri Carlson 2010: FinnWordNet – WordNet på finska via översättning. I: *LexicoNordica 17*: 119–140.

Morris, Jane & Graeme Hirst 2004: Non-classical lexical semantic relations. I: *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*. Boston, Massachusetts: ACL, 46–51.

Murphy, M. Lynne 2003: *Semantic relations and the lexicon*. Cambridge: Cambridge University Press.

Palmer, Martha, Hoa Trang Dang & Christiane Fellbaum 2006: Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering* 13(2): 137–163.

Pedersen, Bolette Sandford, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen & Henrik Lorentzen 2009: DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. I: *Language Resources & Evaluation* 43:269–299.

Piasecki, Maciej, Stanisław Szpakowicz & Bartosz Broda 2009: *A wordnet from the ground up*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Svensén, Bo 2010: Liktydingar med anor. Om den latinska synonymlexikografins utlöpare i Sverige. I: K. Jóhannesson et al. (red.): *Bo 65. Festskrift till Bo Ralph*. Göteborg: Meijerbergs arkiv för svensk ordforskning, 372–382.

Internethänvisningar

DICTIONARY.COM 2011 = http://dictionary.reference.com (juni 2011)

GWN 2011 = http://www.globalwordnet.org (juni 2011)
PWN 2011 = http://wordnet.princeton.edu (juni 2011)
SALDO 2011 = http://spraakbanken.gu.se/saldo (juni 2011)
SB 2011 = http://spraakbanken.gu.se (juni 2011)
SBLEX 2011 = http://spraakbanken.gu.se/sblex (juni 2011)
SYNLEX 2011 = http://folkets2.nada.kth.se/synlex.html (juni 2011)
WIKTIONARY 2011 = http://sv.wiktionary.org (juni 2011)

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

Markus Forsberg
forskare
Språkbanken
Institutionen för svenska språket
Göteborgs universitet
Box 200
SE-405 30 Göteborg
markus.forsberg@gu.se