NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Automatisk identifiering av semantisk föränd - ring med hjälp av lexikala distributionella mönster

Forfatter: Karin Cavallin

Kilde: Nordiske Studier i Leksikografi 12, 2013, s. 106-120
Rapport fra Konferanse om leksikografi i Norden, Oslo 13.-16. august 2013

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

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

Automatisk identifiering av semantisk förändring med hjälp av lexikala distributionella mönster

Karin Cavallin

By looking at word relations in corpora from different time periods we aim to automatically find candidates of semantic variation. We start with the verbal predicate, nominal object relation. This is a syntactic relation that also has an obvious semantic relation. By using techniques from collocational analysis and other statistical measures, we convey differences in distribution that can be indications of semantic variation.

1. Introduktion

Syftet med föreliggande studie är att automatiskt identifiera semantisk förändring. Arbetet ska resultera i listor med ord som är goda kandidater för semantisk variation. Dessa listor kan underlätta för lexikografen då ett första urval görs, men det är sedan upp till lexikografen att bedöma och analysera kandidaterna. Med semantisk förändring avser vi *utvidgad* och *inskränkt* betydelse. Det vill säga ord som har en ökad eller minskad möjlig användning, oavsett av vilken anledning. Semantisk förändring kan ha flera olika skäl som exempelvis

metaforbildning och tabu, men vi är här endast intresserade av sådant som torde kunna mätas. Även *neologismer* och *arkaismer* kommer att ingå i denna studie, men mer som en bonus, då de kan ses som att dra en utvidgad eller inskränkt betydelse till sin spets. Dessutom har det visat sig att det som distributionellt beter sig som en arkaism eller neologism i många fall i själva verket är en inskränkt eller utvidgad betydelse.

Semantisk förändring har tills nyligen varit styvmoderligt behandlat vad gäller komputationella metoder. De befintliga ansatser som har applicerats på semantisk förändring har tillämpat metoder som tidigare använts inom komputationell semantik som exempelvis *word sense induction* (Lau et al., 2012) eller *vector space modelling* (Gulordava och Baroni, 2011). Vektormodeller och induktionsmodeller kräver stora lexikala resurser. Samtliga ansatser har hittills baserats på engelska. För engelska förekommer det förhållandevis stora historiska textresurser, men detta är få språk förunnat.

Ansatsen här är för svenska. Som huvudmaterial används Litteraturbankens copyrightfria texter från 1800-talet. Detta material jämförs dels med Parolekorpusen, dels med SUC-romanerna. Även semantiska förändringar i mer modern tid eftersöks. För dessa studier används pressmaterialet Press65 och Press95. Samtliga korpora distribueras av Språkbanken (spraakbanken.gu.se).

SUC-romanerna är en korpus som består av 58 skönlitterära böcker utgivna 1990-1994. Press65 och Press95 är korpora som består av pressmaterial. Parolekorpusen är en någorlunda balanserad korpus. Vi har dock valt att endast använda det material som är skönlitteratur, och viss presstext.

För att kringgå behovet av enorma korpora har vi valt en syntaktisk relation som även har en stark semantisk koppling:

relationen mellan predikat (verb) och deras nominala argument. För att konformera dessa data används endast nominalfrasens huvudord. Dessa så kallade *verb-objektpar* samlas i *lexikala mängder* (lexical sets). De lexikala mängderna används för att jämföra distributionella mönster.

Varje enskilt verb-objektpar tilldelas ett *log-likelihood*-värde (Dunning, 1993) som visar hur starkt associerat ett verb-objektpar är i jämförelse med de andra verb-objektparen. Detta används för att rangordna och sedan jämföra rangordningen i de lexikala mängderna.

Man kan även titta på skillnaden i hur många *olika* verbtyper ett substantiv förekommer som objekt till, respektive hur många olika substantivtyper verben väljer. Med typer avses distinkta varianter av verb, respektive substantiv, och hänsyn tas inte till frekvens.

Observationerna av distributionella skillnader kan man utnyttja och kvantifiera, och därigenom få listor på ord som är kandidater för semantisk förändring.

Den metod vi utvecklar här är i princip språkoberoende, men kräver att språket i fråga ska ha urskiljbara satsdelar, som utöver den syntaktiska relationen även har ett semantiskt beroende mellan sig.

Upplägget för denna artikel kommer vara att vi presenterar våra korpora, och sedan hur våra data ser ut. Därpå visar vi vilka metoder vi använder, följt av de resultat man får ut. Slutligen ges en sammanfattande diskussion av resultaten.

2. Korpora

De korpora vi använder oss av kommer alla från den svenska Språkbanken (spraakbanken.gu.se). Språkbanken tillhandahåller en stor mängd korpora. För detta projekt har fem korpora använts.

Parolekorpusen, SUC-romanerna, Press65 och Press95 har använts utan större manuell insats. Bortsett från att vi för Parolekorpusen har valt ut en delmängd av den. Vill man inte använda det givna gränssnittet Korp (http://spraakbanken.gu.se/korp/) kan man själv ladda ner korpusarna från Språkbankens hemsida (http://spraakbanken.gu.se/swe/resurser/corpus) och även delvis välja ut specifika delar ur korpusarna som exempelvis litterärt material, eller korpustexter från specifika datum. Den femte korpusen som använts, Litteraturbanken, har vi fått hantera annorlunda.

> Litteraturbankens syfte är att vara en fri kulturhistorisk och litterär resurs för forskning, undervisning och folkbildning. Den är till för alla: forskare, lärare, studerande och litterärt allmänintresserade. Huvuduppgiften är att samla in och digitalisera skönlitteratur och viktigare humaniora samt tillgängliggöra materialet på sådant sätt att det blir möjligt för användare att arbeta med det. (http://litteraturbanken.se)

Numera är Litteraturbanken sökbar under det gängse Korpgränssnittet (http://spraakbanken.gu.se/korp/), men har tidigare inte distribuerats i samma form som övriga. Vi har därför i detta arbete en version av Litteraturbanken som skiljer

sig åt från den distribution som nu finns tillgänglig i Korp.

Litteraturbanken innehåller en stor mängd historiskt material, från så tidigt som 1200-talet, till material skrivet under senare delen av 1900-talet. För denna studie har det material som är från 1800-talet valts ut. Detta gjordes för att ha en någorlunda avgränsad epok att jämföra med, samt för att de språkteknologiska verktyg som behövs för att göra Litteraturbanken tillgänglig för språkteknologiska studier skulle kunna vara tillämpbara. De flesta språkteknologiska verktyg är skapade för att användas på modernt språk.

För att Litteraturbanken (med Litteraturbanken menas nu subkorpusen med 1800-talsmaterial) skulle vara jämförbar med de övriga korpora från Språkbanken användes samma, eller snarlika, verktyg för ordklasstaggning och parsning. Parsningen gjordes med MaltParser (Nivre et al., 2006) som är samma som används inom Språkbanken. För ordklasstaggningen var vi tvungna att frångå HunPos som används inom Språkbanken och istället förlita oss på Trigrams and Tags tagger (T'n'T) (Brants, 2000). HunPos är *open source* och en reimplementation av TnT, som primärt endast skiljer sig åt vid hanteringen av okända ord, där T'n'T presterar bättre (Halaczy et al. 2007). Historiskt material har givetvis en större mängd, för taggaren, okända ord än ett modernare material, då taggaren är tränad på modernt material. För lemmatisering har vi fått materialet delvis lemmatiserat med hjälp av Språkbankens resurser, men även fått komplettera en stor mängd olemmatiserade ord manuellt.[1]

Storleken på korpusarna skiljer sig åt. Istället för att sampla korpusarna i lika storlek har vi valt att utvinna så mycket infor-

1. Det manuellt lemmatiserade materialet är unikt för detta projekt, men kan komma att infogas i Språkbankens distribution av Litteraturbanken framöver.

mation som möjligt ur de korpora vi har till vårt förfogande. Därför har vi normaliserat materialet. För Parolekorpusen har vi emellertid valt ut de delkorpusar som innehåller så litteraturlikt material som möjligt.

2.1. Verb-objektpar och lexikala mängder

En viktig ansats i detta projekt är att använda den uppenbara semantiska information som redan finns mellan vissa satsdelar. Idag utförs den mesta språkteknologin med lexikalsemantisk inriktning med hjälp av metoder som inte har någon lingvistisk grund (men som trots avsaknaden av lingvistisk grund ger ganska goda resultat). I denna studie har vi valt att utnyttja den semantiska relationen mellan verb (verbalt predikat) och objekt (det nominala huvudet i objektsfrasen). Vi kommer hädanefter kalla dessa för verb och objekt, alternativt verb-objektpar. Skillnaden i distributionen mellan verb och objekt mellan olika tidsperioder torde visa på semantiska förändringar.

Verb-objektparen kan förhållandevis enkelt extraheras ur korporan då korporan är parsade, det vill säga uppmärkta med satsdelsinformation.

Verb-objektparen samlar vi ihop, uppdelade på ursprung från respektive korpora. Därefter hanterar vi dem på olika sätt. Vi skapar lexikala mängder, vilket innebär att vi sorterar verben utifrån vilka objekt verben samförekommer med, och objekten utifrån vilka verb de samförekommer med. Nedan visas exempel på lexikala mängder, som utgår från en konstruerad text.

> Det är sommar. Lisa är svettig. Hon köper en glass. Hon dricker även en svalkande läsk. Men den var inte särskilt god så hon slängde läsken. I eftermiddag tar hon bilen till landet. Hennes bror har köpt ett hus. Hon har köpt en ryggsäck till resan. Hon köpte även vin och tre liter vatten och en glass till sin bror. Ikväll ska de dricka vatten, vin och öl. Barnen får dricka mjölk. Hunden tog en promenad och kastade vatten innan de begav sig iväg.

Utifrån texten ovan extraheras lemmatiserade verb och objekt. Den extraherade datamängden redovisas nedan:

> {köpa glass, dricka läsk, slänga läsk, ta bil, köpa hus, köpa ryggsäck, köpa vin, köpa vatten, köpa glass, dricka vatten, dricka vin, dricka öl, dricka mjölk, ta promenad, kasta vatten}

Från datamängden konstrueras sedan lexikala mängder som konstitueras, antingen utifrån det gemensamma verbet för objekten, eller utifrån det gemensamma objektet för verben.

> köpa: glass, ryggsäck, hus, vatten
> dricka: läsk, öl, vatten, vin, mjölk
> ta: bil, promenad
> glass: köpa
> läsk: dricka, slänga
> öl: dricka
> vatten: dricka, kasta
> bil: ta
> promenad: ta

Från och med nu och fortlöpande i texten kommer vi inte längre att förhålla oss till korpusar, utan endast till datamängden som består av verb-objektpar. Datamängden kan vi även kalla lexikala mängder, särskilt då vi syftar på specifika lexikala mängder, men även som ett överbegrepp för innehållet i datamängden.

3. Metoder

Det finns ett antal grundantaganden som vi har gjort inför datahanteringen:

1. Det finns en semantisk, såväl som en syntaktisk relation, mellan ett verb och dess objekt
2. Skillnad i frekvens kan vara en första indikator på att ett ords betydelse skiljer sig åt mellan olika korpora.
3. Närheten, associationen, mellan verbet och objektet kan mätas med hjälp av metoder använda inom kollokationsforskning.
4. Rangordningen inom två motsvarande lexikala mängder kan påvisa en semantisk variation.
5. Antalet olika objekt- eller verbtyper ett verb, respektive objekt, förekommer med, kan vara en indikation på semantisk variation.

En viktig sak att peka på är att våra data är normaliserade. Detta görs för att de lexikala mängderna och datamängderna trots sin

olika storlek ändå ska vara jämförbara. Punkt 2 och 5 förhåller sig till normaliserade värden. I punkt 2 är normaliseringen gjord med avseende på antal *tokens* i den specifika lexikala mängden över antalet tokens i datamängden den lexikala mängden utgår från. I punkt 5 normaliserar vi antal *typer* i en specifik lexikal mängd över antalet tokens i den datamängd den lexikala mängder utgår från.

För att få fram kandidater för semantisk variation väljer vi att kombinera dessa antaganden och vaskar fram de verb eller objekt som uppfyller samtliga kriterier i punkt 2-5.

Punkt 1 är inte en metod, utan snarare en teoretisk grund för att övriga kriterier/punkter ska kunna vara behjälpliga för att hitta semantisk variation.

Punkt 2. Vi extraherar de verb eller substantiv som skiljer sig åt med avseende på frekvens, dvs. att ett ord förekommer betydligt fler gånger i en datamängd än i den andra

Det är viktigt att poängtera att vi här endast tittar på två datamängder i taget. Vi jämför Litteraturbanken med Parole, Litteraturbanken med Stor-SUC, samt Press65 och Press95 med varandra.

Punkt 3. Med hjälp av ett log-likelihood värde (Dunning, 1993) mäts associativiteten mellan ett verb och objekt och deras relation till övriga verb och objekt i datamängden. Par som är mycket närmare associerade i en datamängd i jämförelse med den andra datamängden ger en indikation på semantisk variation.

Punkt 4. Spearman ranking coefficent (Spearman, 1904) är ett mått som här mäter skillnader i rangordningen av verb-objektparen mellan två motsvarande lexikala mängder. Spearman ranking coefficient avger ett värde mellan -1 och 1. Ju lägre värdet är, desto mer olika är de motsvarande lexikala mängderna. De lexikala mängder som skiljer sig åt mest ger följaktiktligen en indikation på semantisk variation.

Punkt 5. En utvidgad eller inskränkt betydelse borde praktiskt påvisas med hjälp av att det förekommer andra typer, eller en större mängd typer i den ena än i den andra datamängden.

4. Resultat

Genom att kombinera metoderna listade ovan kan man identifiera distributionell variation som kan vara indikation på semantisk variation. Tanken är att dylika listor, som i tabellen nedan (Tabell 1), kan användas av någon som vill kartlägga förändringar eller variationer inom förslagsvis en specifik domän, jämföra terminologin i två närliggande discipliner, eller för att hitta generella förändringar.

I tabellen nedan väljer vi att kalla kategorierna för «Utvidgad bet.», «Inskränkt bet», «Arkaism» och «Neologism». Orden är endast *kandidater* för dessa kategorier, det är alltså en *potentiell* utvidgning, arkaism och så vidare. Vi kommer här att redovisa en delmängd av de kandidater vi fått från en studie av Litteraturbanken och Stor-SUC och föra ett kort resonemang om intressanta fall i tabellen.

I punktlistan nedan sammanfattar vi de kriterier som spelar in för att få fram respektive variationstyp.

- «Utvidgad betydelse» ← utökad typ- och tokenförekomst, samt starkt varierad rangordning
- «Inskränkt betydelse» ← minskad typ- och tokenförekomst, samt starkt varierad rangordning.

- «Arkaism» ← minimal typ- och tokenförekomst i den ena datamängden (den yngre) men inte i den andra (den äldre).
- «Neologism» ← minimal typ- och tokenförekomst i den ena datamängden (den äldre) men inte i den andra (den yngre).

De ord som är homografer över ordklasser är markerade med frågetecken, «?», ord som är uppenbara fel vad gäller taggning är markerade med asterisk, «*».

Utvidgad bet.	**Inskränkt bet.**	**Arkaism**	**Neologism**
?rätt	?för	konung	?far
råd	hjärta	anlete	statsminister
problem	natur	hop	skit
nummer	hustru	*själv	*sig
kurs	?vår	grevinna	Kollega
	syster	sällhet	Grej
	händelse	majestät	bomb
	?karl	Fröjd	statsråd
	?tro	tycke	värdering
	karaktär	skald	styck
	vagn	?före	relation

Tabell 1: Tabell över de ord som skiljer sig åt mest i Litteraturbanken och Stor-SUC m.a.p. de kriterier som beaktats.

De ord som står i kolumnen «Utvidgad bet.» är nog svårast att se om de är dugliga kandidater. Tidigare studier (Cavallin, 2012) har visat att *problem* är ett ord som har fått utvidgad betydelse, från ett mer abstrakt problem, till mer vardagliga och praktiska bekymmer. Övriga ord kräver mer ingående studier för att avgöra om de är dugliga kandidater. Under kolumnen «Inskränkt bet.» hittar vi fler kandidater som vi, med en generell språkkänsla för svenska språket, kan se som eventuellt goda kandidater. *Hustru* och *karl* är ord som drar åt det arkaiska, och i vissa grupper även behäftade med en anstrykning av negativ laddning. *Tro* är kanske inte förändrat i betydelse, men ett betydligt mindre använt ord i det numera sekulariserade Sverige. De ord som är kandidater under «Arkaism» är samtliga, utom möjligen *tycke* och *hop*, ord som har en arkaisk klang. Under «Neologism» ser vi att det knappast är helt nya ord, men det är ord som betecknar modernare företeelser.

Det måste nämnas att det är oändligt svårt att dra gränsen mellan samhällelig och semantisk förändring. De är givetvis inte oberoende av varandra, men något kan helt enkelt vara «inne», utan att ha förändrat betydelse. Aktuella och dramatiska händelser kan exempelvis påverka framförallt tidningstext, utan att ordet i sig har ändrat betydelse. Som exempel brukar *tsunami* lyftas fram. Få svenskar visste vad ordet betydde, eller ens att det förekom i svenskan förrän den stora tsunamin 2004, där bland tusentals andra, ett stort antal svenska turister förolyckades.

Utöver de mer lingvistiska vinningarna kan utdata i form baserad på syntaktisk och semantisk relation bidra till att exempelvis korrigera taggnings- och parsningsverktyg, och därigenom förbättra materialen.

De lexikala mängderna fungerar också som en minimal kontext, som kan göra det enklare att snabbt få överblick över vissa lingvistiska fenomen.

Dock måste alla resultat kritiskt granskas av en expert. Ansatsen här ger en första och ganska tydlig bild av vilka ord som uppvisar semantisk variation i de korpora man är intresserad av att studera.

Litteratur

Länkar

Litteraturbanken: http://litteraturbanken.se/
Språkbankens hemsida för nedladdningsbara resurser: http://spraakbanken.gu.se/swe/resurser/corpus
Korp: http://spraakbanken.gu.se/korp/

Ordböcker

SAOL10 = *Svenska akademiens ordlista över svenska språket* (10:e utg.) (1976). Svenska Akademien.
SAOL13 = *Svenska akademiens ordlista över svenska språket* (13:e utg.) (2006). Svenska Akademien.

Annan litteratur

Brants, Thorsten (2000): TnT–A Statistical Part-of-Speech Tagger. I: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*. Seattle, WA, 224-231.

Cavallin, Karin (2012): Automatic extraction of potential examples of semantic change using lexical sets. I: Jeremy Jancsary (red.): *Proceedings of KONVENS 2012 LThist 2012 Workshop.* Wien: ÖGAI,. 370–377.

Dunning, T. (1993): Accurate methods for the statistics of surprise and coincidence. I: *Computational Linguistics 19 (1)*, 61–74.

Gulordava, Kristina & Marco Baroni (2011): A distributional similarity approach to the detection of semantic change in the google books ngram corpus. I: S. Pado & Y. Peirsman (red.) *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, July 2011*. Edinburgh: Association for Computational Linguistics, 67–71.

Halácsy, Péter, András Kornai & Csaba Oravecz (2007): Hunpos: an open source trigram tagger. I: *Proceedings of the 45th annual meeting of the ACL on Interactive Poster and Demonstration sessions, ACL '07*. Stroudsburg, PA: Association for Computational Linguistics, 209–212.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman, & Timothy Baldwin (2012): Word Sense Induction for Novel Sense Detection. I: M. Butt et. al. (red.) *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, *Avignon, France, April 2012*:. Stroudsburg, PA: Association for Computational Linguistics, 591–601.

Nivre, J., J. Hall & J. Nilsson (2006): MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. I: N. Calzolari et. al. (red.) *Proceedings of the fifth international conference on Language Resources and Evaluation. (LREC2006), May 24-26, 2006, Genoa, Italy*, 2216–2219.

Spearman, Charles (1904): The Proof and Measurement of

Association between Two Things. I: *The American Journal of Psychology 15 (1)(January*), 72–101.

Karin Cavallin
forskarstuderande, fil.mag.
Institutionen för filosofi, lingvistik och vetenskapsteori
Göteborgs universitet
Box 200
405 30 Göteborg
karin.cavallin@gu.se