Automatic extraction of potential examples of semantic change using lexical sets

Karin Cavallin
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Sweden
karin.cavallin@gu.se

Abstract

This paper describes ongoing work on automatically finding candidates for semantic change by comparing two corpora from different time periods. Semantic change is viewed in terms of distributional difference with a computational and linguistically motivated approach. The data is parsed, lemmatized and part of speech information is added. In distributional semantics, meaning is characterized with respect to the context. This idea is developed from Firth (1957) and is formulated according to 'the distributional hypothesis' of Harris (1968). A method is developed to describe distributional behaviour in order to track semantic change over time. We will explore statistically ranked lists of *verbal predicate - nominal object* constructions and examine differences at the level of word types.

1 Introduction

When I asked a lexicographer how he finds *semantic change* the answer was "I read a lot". This method, however pleasant, might not be the most efficient way, and we here propose a way of extracting candidates for semantic change automatically, in the hope that this gives the lexicographers more time to do the proper analysis of the candidates instead of time-consuming reading. The main idea is to make a quantitative study in order to automatically find candidates for semantic change in two corpora, one of 19th century data and one of data from 1990's. In this exploratory stage of the project, we work with transitive verb predicates and nominal objects.

A distributional standpoint is adopted, in that meaning is characterized in terms of the context in which words occur. The distributional hypothesis is attributed to Harris (1968) but the most famous quote representing this view is by Firth (1957): "You shall know a word by the company it keeps"[1]. The distributional hypothesis is not uncontroversial, but for computational means this assumption has proven fruitful in several tasks.

The data used for this project consisted of 19th century Swedish literary texts from *Litteraturbanken* (the Swedish Literature Bank) (LB) and parts of the Swedish Parole corpus (PA) with text from the 1990's.

We expect that a comparison of the ranks and other differences in distribution, both on frequency level and type level, will yield information that will pick out candidates for semantic change. We use the terms "type" and "token" as it is standard in corpus linguistics. Whereas *token* refers to number of word occurrences, *type* refers to the (orthographic) representation of a word group.

1.1 Related work

The field of semantic change seems to have received little attention in natural language processing (NLP). There has been a small amount of research in the last few years. Sagi et al. (2009) adapt latent semantic analysis and present a way of detecting semantic change based on a semantic *density* measure. The density is conveyed by the cohesiveness of the vectors. Cook and Hirst (2011) focus on finding amelioration and pejo-

[1] A thorough survey of distributional semantics is found in Sahlgren (2006) and Sahlgren (2008).

ration. Gulordava and Baroni (2011), "address the task of automatic detection of the semantic change of words in quantitative way" focusing on detecting semantic change rather than specifically widening, narrowing or emotive change. Rohrdantz et al. (2011) "presents a new approach to detecting and tracking changes in word meaning by visually modelling and representing diachronic development in word contexts"[p.305].

The GoogleNgram-viewer (Michel et al., 2011), provides a quick and easy way of detecting changes in ngram frequencies, which can be interpreted for different aspects of cultural change. Lau et al. (2012) "apply topic modelling to automatically induce word senses of a target word, and demonstrate that [their] word sense induction method can be used to automatically detect words with emergent novel senses, as well as token occurrences of those senses."[p.591].

Hilpert (2012) works with diachronic collostructional analysis, which is mainly about semantic change in grammatical constructions. So far there is no consensus or standard way of approaching semantic change from a more quantitive perspective.

1.2 Lexical sets

We work with syntactically motivated collocational pairs, in this case verbal predicates and the head noun of object arguments, from here on 'verb-object pairs', we collect these together in *lexical sets*. Any set of lexical units that share a common feature can constitute a lexical set, be it phonological, ontological, orthographical, etcetera. Here, a lexical set can be a *verbal lexical set*, the verbs that occur as governing verb to a given nominal argument, or a *nominal lexical set*, the nouns that occur as argument to a given verb[2]. The nominal and verbal lexical sets are arranged as ranked lists according to an association measure.

1.3 Log-likelihood

The ranking is based on a log-likelihood (Dunning, 1993) count and the verb-object pairs with stronger association are ranked higher. The log-likelihood implementation is taken from the Ngram Statistic Package (Banerjee and Pedersen, 2003). "The log-likelihood ratio measures the deviation between the observed data and what would be expected if <word1> and <word2> were independent. The higher the score, the less evidence there is in favor of concluding that the words are independent."[3]. The log-likelihood measure is applied to the extracted verb-object pairs in the corpora, thus providing us with a ranking such that the higher the ranking, the less likely it is that the words are independent within the respective pair, that is they have a stronger association with each other. This measure has primarily been used for collocation extraction and therefore seems appropriate to verb-object pairs.

The data here is summarized in ranked lists from the perspective of the predicate or argument, respectively. Senses will be analysed through *frequency* and a log-likelihood based *ranking* and also the difference in number of different *types* a word co-occurs with in its lexical set.

2 Preparing the data

2.1 The data

In order to track semantic change over time there is a need for corpora containing material from different time periods. For modern Swedish we use a selected subset of the Parole corpus (Språkbanken, 2011). The subcorpus we use contains novels, magazines, press releases and newspaper texts from the 1990s. For the older Swedish we use the Swedish Literature Bank (Litteraturbanken, 2010), a resource not adapted for NLP purposes, but intended for people with literary interests. The Swedish Literature Bank carries material from as early as the 13th century, but for the present study only the 19th century texts have been tagged, parsed and lemmatized.

The subcorpus of the Swedish Literature Bank we are building amounts to approximately 10 million word tokens. Of these approximately 2 million are tagged as nouns, and 1.5 million as verbs[4]. The subcorpus of Parole amounts to approximately 8 million tokens, whereof 1.4 million

[2]The definition and the ranking of *nominal* and *verbal* lexical set follow the work of Jezek and Lenci (2007).

[3]`http://search.cpan.org/~tpederse/Text-NSP-1.25/lib/Text/NSP/Measures/2D/MI/ll.pm`, 2012-08-16

[4]Accuracy has not been estimated yet.

are words tagged as nouns, and 1.3 million tagged as verbs.

Table 1: No. of nouns, verbs and extracted verb-object pairs in the datasets.

	LB	PA
Word tokens	10M	8M
Noun tokens	2M	1.4M
Verb tokens	1.5M	1.3M
Extracted VO pairs tok	290.878	482.221
Extracted VO pairs typ	157.869	97.077

We see that we have been able to extract more VO pairs from the PA corpus, despite the fact that it is slightly smaller. However, the number of different pair-types also differs within the datasets. There is, for example, a much greater difference between the verb-object pair tokens than the verb-object pair types in the PA data set. This suggests that looking at types is a more promising starting point than raw frequency since that might be less sensitive to, for instance, over-representation.

2.2 PoS-tagger

The PoS-tagger used for the Swedish Literature Bank is Trigrams'n'Taggers, TnT (Brants, 1998), since its output is compatible with the MaltParser (see section 2.3). Good results are also attested in tagging texts containing many misspellings[5] such as those of primary school students.[6] Treating 19th century spellings as if they were misspellings is one heuristic way of addressing the problem of tagging 19th century Swedish. During the manual lemmatization (described in section 2.4), the PoS errors detected were omitted from the final data set.

2.3 Parser

One of the most important features in pursuing this sense tracking is to ensure that the corpus is parsed in order to identify predicates and objects. A parser freely available and widely used in the NLP community is the MaltParser (Nivre and Hall, 2005). The MaltParser is a system for data-driven dependency parsing. We have used a pre-trained Swedish model available from the MaltParser distribution. This model is of course trained on modern Swedish, which gives rise to noise in non-modern data, but we hope this is insignificant given the amount of data. In a recent paper, Pettersson et al. (2012) attempt to improve parsing by normalising their data from 1550-1880 (i.e. mostly by normalization spelling before tagging and parsing). We hope to take advantage of this result in future work.

2.4 Lemmatization

Some of the material was lemmatized automatically. The automatic lemmatization works as follows: There are two linked morphologies. One is originally a 19th century dictionary, Dalin (Dalin, 1850–1853), now enhanced with a full-form morphology[7]. The contemporary morphology is SALDO (Borin and Forsberg, 2008), an among other things full-form morphology lexicon. Both were developed at Språkbanken at Gothenburg University. Given the PoS information, which reduces ambiguity, the base form is extracted, and via the linking to SALDO, a contemporary base form (lemma) is extracted.[8]

However, in order to improve coverage, we performed a manual lemmatization on all the unlemmatized verbs and nouns in the LB corpus. The lemmatization is on the word form level and semantic ambiguities are not resolved. This is partly for practical reasons and partly to make the material as unbiased as possible with regard to sense, and also to avoid discussion of how fine-grained distinctions should be.

3 Method and Theory

The main data is basically a data set with information regarding ranked frequency occurrence from different perspectives and values computed on these frequencies by different statistical measures for the two data sets of extracted verb-object pairs of the Swedish Literature Bank and the Parole corpus.

The ranking can be made in different ways. By not ranking merely according to raw frequency,

[5] Personal communication, Sofie Johansson Kokkinakis.

[6] As far as I know there is currently no account of why the TnT works well with misspellings.

[7] http://spraakbanken.gu.se/eng/resource/dalin

[8] This was carried out by Markus Forsberg, employing a similar lemmatization approach to that used, but not presented, in Borin et al. (2010) and Borin et al. (2011).

but adding a statistical measure to the data set (here a log likelihood count), we get a more reliable pattern where the most strongly associated pairs are ranked higher. This is a common approach for collocational extraction (Manning and Schütze, 1999) which this lexical set extraction shares similarities with. This data provides a basis for extracting viable candidates for semantic change.

It would be too ambitious to attempt to find *reasons* for semantic change automatically. What we should be able to find is the *consequences of semantic change*. We follow the distinction of consequences of semantic change of *widening*, *narrowing*, *amelioration* and *pejoration*, discussed by Bloomfield (1933). Whether a detected change is a widening, narrowing or a difference in emotive value is of course not detected by a difference in number, but needs manual lexical inspection to be determined. However, an increase or decrease in different types (as in item 3 below) should be a good indicator of a widening or narrowing, respectively. The latter two are not possible to perform computationally without a proper ontology. We want to consider the following aspects:

1. Increased or decreased raw frequency of a given *verb*, *object* or *verb-object pair*.
2. A higher or lower rank, given a statistical measure of a verb-object pair.[9]
3. Increased or decreased number of different types for a given verb or object.
4. Difference in the semantic type or class of the words in a lexical set.
5. A summarized difference of the semantic types of the respective lexical sets.

Differences in frequency between the two corpora is a first, but crude, indicator of semantic change. An ontologically motivated analysis would provide greater insights, but there is no full-coverage ontology available for Swedish. A more useful and theoretically motivated strategy is looking at the number of different types as shown in section 5.

[9]Here log-likelihood has been used, other association measures can also be applied.

4 Preliminary results

4.1 Lexical sets as means to finding semantic change

First attempts (Cavallin, 2012) showed distributional differences for what would be a widening. The noun "kontakt", 'contact', must have undergone a widening, from merely a closeness between surfaces, into a connection between people. This is manifested as an increase of raw frequency (1878 occurrences in the later period versus 9 occurrences in the earlier), and also a higher rank in the lexical sets extracted from the Parole corpus (140 versus 7666). With respect to difference in type frequency we find a striking difference of 68 different verb (types) solely occurring in Parole (PA), five only occurring in the Literature Bank (LB), and two occurring in both. Thus information about raw frequency, ranking and type difference give us strong indications that there has been a change. If we look further at the lexical set of "kontakt" in Table 2, we also find that there are many words in the top fifteen that refer to "kontakt" in the sense of social connection, rather than the contact of surfaces or an electric plug. The last items on the list, which only occur in the Literature Bank seemingly mainly concern physical contact.

5 Type level differences as means to finding semantic change

In this next step we find a way of comparing lexical sets from a *type* perspective, and make this information give us a set of viable candidates for semantic change.

By extracting the elements in the datasets (mentioned in section 2) that differs the most, we get an overview of candidates for semantic change. The elements in the respective dataset that differ the most are computed by first normalizing the type counts, i.e. the number of different types each given verb/object governs or is argument to (this does not take frequency into consideration). By dividing the given count by the maximal type count of each set the values are normalized and hence comparable by taking the difference between the value of the PA data and the LB data. The greater difference on type level between a given word in the two datasets the higher the po-

Table 2: Lexical set of "kontakt".

Trans	Verb	PaRank	LbRank
take	ta	140	-
have	ha	686	-
hold	hålla	1003	-
give	ge	1871	-
tie	knyta	2154	-
establish	etablera	2157	-
loose	förlora	2468	19175
find	finna	3684	-
come	komma	7134	-
loose	tappa	10118	-
miss	sakna	10806	-
avoid	undvika	10873	-
re-establish	återknyta	11533	-
get	få	11933	-
convey	förmedla	12332	-
...	...	...	...
seek	söka	13322	21363
...	...	...	...
connect	koppla	-	6423
screw	skruva	-	6907
maintain	bibehålla	-	13852
see	se	-	39260

sition in the table (see Table 3). The word type displaying the maximal difference is then ranked highest under *DiffTyp*. *DiffToken* is the difference of the normalized frequency, and does not constitute the rank in the given table. *LbSum* and *PaSum* refer to the unnormalized raw frequency in the data sets. *LbTyp* and *PaTyp* refer to the unnormalized number of types in the datasets.

A high position can indicate a widening, whereas a low position indicates a narrowing.[10] Words in the middle of the type ranking should then not have experienced any major changes. Where to draw the line between what is a significant difference and thus a candidate for semantic change is not straightforward and left for future discussion.

The top word in Table 3, of words displaying the words that differ most regarding the number of types they occur with, is "procent" *percent*. Looking more thoroughly at the lexical sets of "procent" we see that the 12 verbs in the LB data are

[10]So far only widening has been explored.

Table 3: Examples of top objects differring in number of verb types.[12]

Transl.	Object	Lb/PaTyp	Lb/PaSum	DiffTyp/Tok
percent	procent	12/155	17/796	0.31/0.03
crown	krona	115/215	238/2611	0.28/0.03
problem	problem	27/140	39/3134	0.26/0.11

fairly vague, or semantically light (such as "kalla" 'call', "ha" 'have' , "göra" 'do', "få" 'get', "ta" 'take'). The 155 different verbs occuring with "procent" in the PA data show many verbs that are semantically rich, for example: "kontrollera" 'control', "äga" 'own', "samla" 'collect', "producera" 'produce' and "utgöra" 'constitute'. It appears that the word "procent" is used in a wider context in PA than in LB, where it occurs primarily in connection with money, a sense attested in the begining of the 18th century. However, the more general notion of "procent" as part of a whole is also attested in the beginning of 18th century (SAOB, 1954)[p.1932ff]. Even though the sense of PART OF A WHOLE is listed in traditional dictionaries, the widening of the *usage* of this interpretation of "procent" is not noted as gaining in the traditional Swedish dictionaries.

"Krona", as a unit of currency, was introduced in 1873, when the Scandinavian coin-union was created.[13] Looking at the lexical set for "krona" ranked according to the difference of the normalized log likelihood values we see a great number of senses among the top ten which refer to MONEY rather than to something WORN ON THE HEAD. This shows us that counting types can point us in the direction of semantic change. However, we must always confirm candidates by manual inspection of the lexical set. This distributional difference for "krona" is definitely due to a non-linguistic historical change which caused the widening. Whereas some semantic change evolves slowly, this newer sense of "krona" has been very actively created and thus easily dated.

"Problem" has a major type and frequency increase. In the early 20th century "problem" is attested in compounds as in "problembarn", 'problem child' (SAOB, 1954)[p.p.1927ff], where the notion is more of a psycho-social problem than

[13]krona. http://www.ne.se/lang/krona/232211, Nationalencyklopedin. Accessed July 13, 2012.

e.g. philosophical and political problems. The more concrete or easier form of problem as 'difficulty' or 'trouble' is more widely used in PA. The widened sense of "problem" has an impact on the frequency and number of types.

We see that words can have kept their original sense into the 20th century, and that even though the subsense has been co-occurring all along, it increases in usage, which we define as a widening of the sense, especially when the increased usage seem to be prevailing.

The approach is a promising contribution towards an automatic prediction of semantic change. It can suggest candidates for widening and narrowing. It would, however, benefit from more accurately parsed and tagged data. By combining the different indicators of semantic change in item 3 we would get even more reliable predictions.

6 Conclusion and future work

Differences in frequency is a first, but crude, indicator of semantic change. A more useful and theoretically motivated approach is looking at the number of different types as shown in the present paper, and assessing the candidates by looking at the appropriate lexical sets. (An ontologically motivated analysis would provide even greater insights, but there is no full-coverage ontology available for Swedish). Semantic change appears to reveal itself distributionally in different ways. Summarizing, it can be *measurably* manifested as:

- an increase in raw frequency (*Tokens*)
- differences in the number of types (*Types*)
- differences in the ranking of the words in the corresponding lexical sets
- differences in the lexical distribution in the corresponding lexical sets

It is important to note that not all distributional differences can be assumed to be semantic change *per se*. One must resort to manual inspection (and theoretical considerations) in order to confirm cases of semantic change. "Krona" is a semantic change brought about by a change in the world. However, there can also be socio-historical changes reflected in language. In our data we have the example of the verb-object pair "läsa-bibel" 'read-bible'. This has a much lower frequency in the modern data in comparison with for instance "läsa-tidning" 'read-news paper'. What is reflected in the data is not that "läsa" has changed, but rather the fact that Sweden has gone through a secularization. At this level of analysis there is no obvious way to distinguish socio-historical change from semantic change by distributional means.[14]

It is important to remember that there are tagging errors where non-nouns (or non-objects) and non-verbs (or non-root predicates) have been tagged as nouns and verbs, especially in the older data. This can wrongly increase the number of different types in the LB data. If we combine the different measurable aspects mentioned above, the errors can hopefully be marginalized awaiting more accurately annotated data.

By comparing (fairly[15]) comparable corpora from different time periods we can be made aware of changes. Given a more fine-graded time distinction of the corpora, we could even attempt tracking where the sense starts being polysemous and where the new sense possibly exceeds the older sense in frequency.

The manually made lemmatization is a valuable resource in enhancing search in the 19th century material in the Swedish Literature Bank, and can be used for building better automatic lemmatization on other older Swedish material.

We would like to compare lexical sets where the given words are within semantically similar domains, which presumably will render further input in the pursuit of semantic change. We are also planning to compare the data taking the corresponding lexical sets and compute which lexical sets are the closest and most distant from each other.

[14] However, an ontological analysis could at least point us in the direction of where the predicate or argument is of a different semantic type. "Krona" would be distinguished as a candidate of semantic change, whereas "läsa" would not, since "bibel" and "tidning" are both PRINTED MATTER, and for instance the verbs in "trycka-krona" 'coin-crown' and "pryda-krona" 'decorate-crown' are ontologically distant.

[15] We are aware of the fact that the difference attested could be a difference only between the corpora, rather than between time periods.

The outcome of the present work is a starting point for automatically detecting semantic change. The approach gives us indications, but there is still a need for manual inspection, which we hope to decrease as resources and methods are refined. This approach is not restricted to Swedish, and would benefit from an attempt in a language with more elaborate language resources.

References

S. Banerjee and T. Pedersen. 2003. The Design, Implementation, and Use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City, February.

L. Bloomfield. 1933. *Language*. Henry Holt, New York.

L. Borin and M. Forsberg. 2008. Saldo 1.0 (svenskt associationslexikon version 2). Språkbanken, Göteborg universitet. `http://spraakbanken.gu.se/eng/saldo/`.

L. Borin, M. Forsberg, and D. Kokkinakis. 2010. Diabase: Towards a diachronic blark in support of historical studies. *LREC2*.

L. Borin, M. Forsberg, and C. Ahlberger. 2011. Semantic search in literature as an e-humanities research tool: Conplisit – consumption patterns and life-style in 19th century swedish literature. In *NEALT Proceedings Series (NODALIDA 2011 Conference Proceedings)*, volume 11, pages 58–65.

T. Brants. 1998. TnT - Statistical Part-of-Speech Tagging. `http://www.coli.uni-sb.de/~thorsten/tnt/`.

K. Cavallin. 2012. Exploring semantic change with lexical sets. In *Proceedings of the 15th EURALEX International Congress*, pages 1018–1022.

P. Cook and G. Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language Information and Computation (PACLIC 25)*, pages 265–274, Singapore, December.

A.F. Dalin. 1850-1855. *Ordbok Öfver svenska språket*, volume I, II. Svenska Akademien, Stockholm.

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

J.R. Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, pages 1–32.

K. Gulordava and M. Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July. Association for Computational Linguistics.

Z. Harris. 1968. *Mathematical Structures of Language*. John Wiley and Son, New York.

M. Hilpert. 2012. Diachronic collostructional analysis: How to use it and how to deal with confounding factors. In K-L. Allan and J.A. Robinson, editors, *Current Methods in Historical Semantics*, pages 133–160. De Gruyter Mouton, Berlin.

E. Jezek and A. Lenci. 2007. When GL meets the corpus: A data-driven investigation of semantic types and coercion phenomena. In *Proceedings of the 4th International Workshop on Generative Approaches to the Lexicon*, Paris.

J H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 591–601, Avignon, France, April.

Litteraturbanken. 2010. `http://litteraturbanken.se/`.

C. Manning and H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA.

J-B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, The Google Books Team, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, January 14.

J. Nivre and J. Hall. 2005. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005)*, pages 137–148, December.

E. Pettersson, B. Megyesi, and J. Nivre. 2012. Parsing the Past - Identification of Verb Constructions in Historical Text. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities.*, Avignon, France.

C. Rohrdantz, A. Hautli, T. Mayer, M. Butt, D.A. Keim, and F. Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 305–310, Stroudsburg, PA, USA. Association for Computational Linguistics.

E. Sagi, S. Kaufmann, and B. Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings*

of the EACL 2009 Workshop on GEMS: GEometrical Models of Natural Language Semantics, pages 104–111, Athens, Greece, March.

M. Sahlgren. 2006. *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* Ph.D. thesis, Stockholm University.

M. Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):1–18.

SAOB. 1954. *Svenska Akademiens Ordbok*. Svenska Akademien, Lund.

Språkbanken. 2011. `http://spraakbanken.gu.se/korp/`.