Exploring semantic change with lexical sets

Karin Cavallin

Keywords: *lexical sets, semantic change, language technology*.

Abstract

Many areas of linguistics which use corpora as their main data have benefited from research in natural language processing, NLP. Apart from a few recent studies such as Sagi et al. (2009), Rohrdantz et al. (2011) and the GoogleNgram-viewer (Michel et al. 2011), the field of semantic change seems to have received little attention in NLP. This paper describes some first steps in viewing semantic change in terms of distributional semantics with a computational and linguistically motivated approach. By parsing, adding lemmatization and part of speech information, a method is developed to describe semantic behavior and to track semantic change over time. 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). Whereas most approaches to statistical semantics uses some kind of vector analysis based on ngrams. Distribution here is presented as the statistically ranked lists of verb-object constructions, that is 'lexical sets'. A lexical set is more focused than ngrams and can be seen as essential minimal co-occurrence information for a given word, which facilitates manual analysis.

1. Introduction

Many areas of linguistics which use corpora as their main data have benefited from research in natural language processing, NLP. Apart from a few recent studies such as Sagi et al. (2009), Rohrdantz et al. (2011) and the GoogleNgram-viewer (Michel et al. 2011), the field of semantic change seems to have received little attention in NLP.

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". (A thorough survey of distributional semantics is found in Sahlgren (2006, 2008)). In this paper we present a method for the investigation of Swedish 19th century data in comparison with Swedish 20th century data via annotation with syntactic information.

To be able to manually inspect the large amounts of data we work with a hypothesis of syntactically motivated minimal concordance co-occurrence pairs, that is in this case verbal predicates and their head noun of their objects[1], from here on 'verb-object pairs'. This is what Jezek and Lenci (2007) refer to as lexical sets. The data is summarized in ranked lists from the perspective of the governing or argument word, respectively. Senses are analyzed through frequency and different statistical rankings, but also how words co-occur distributively in their respective lexical sets. For this study we have chosen 'the log-likelihood measure' from the Ngram Statistic Package (Banerjee and Pedersen, 2003) as the statistical measure. This is calculated on the extracted verb-object pairs in the corpora. This provides us with a ranking where the higher the ranking, the less likely it is that the words are independent within the respective pair, that is they have a strong association with each other. By using lexical sets we can find changes using semi-automatic means by being aware of differences in distribution, frequency and ranking.

To exemplify, the first eight items of the lexical set of *läsa* 'to read' and *bok* 'book' from the 20th century is presented in Table 1. Except for information regarding their place in the overall ranking of all verb-object pairs in the 20th century data, the lexical sets are stripped of

their statistical information for clarity.

Table 1. Example of a nominal and a verbal lexical set.

Rank	Verb-Object	Translation	Rank	Verb-Object	Translation
87	~ bok	~ book	87	läsa ~	read ~
199	~ tidning	~ newspaper	251	skriva ~	write ~
215	~ brev	~ letter	807	ha ~	have ~
328	~ dikt	~ poem	1435	slå ~	hit ~
631	~ juridik	~ law[2]	1520	spela ~	play ~
804	~ läxa	~ homework	1539	få ~	receive ~
1102	~ roll	~ part (in a play)	1629	ge ~	give ~
1107	~ rad	~ row	3477	utge ~	publish ~

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. We use the Swedish Literature Bank (Litteraturbanken, 2010), a resource intended for lay people and professionals with literary interests. We also use a selected subset which consists of novels, magazines, press releases and newspaper texts of the Parole corpus (Språkbanken, 2011). The Swedish Literature Bank carries material from as early as the 13th century, but for present study we have tagged, parsed and lemmatized the 19th century data. This is chosen as a point where a modern tagger and parser can have a reasonable chance of producing acceptable analyses. Furthermore, much of the older material is scanned text that would have to undergo OCR and would thus provide an additional level of complication.[3]

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[4], and 1.5 million as verbs. The subcorpus of Parole amounts to approximately 8 million tokens, whereof 1.4 million are words tagged as nouns, and 1.3 million tagged as verbs.

2.1.1 *PoS-tagger*. The PoS-tagger used for the Swedish Literature Bank is the Trigrams'n'Taggers, TnT (Brants, 1998), for which good results are attested in tagging texts containing many misspellings such as those written primary school students.[5] Viewing 19th century spellings as misspellings is one heuristic way of addressing the problem of tagging 19th century Swedish.

2.1.2 *Parser*. One of the most important features in pursuing the sense tracking of minimal concordance co-occurrence pairs 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 noise in non-modern data, but we hope this to be insignificant given the amount of data.

2.1.3 *Lemmatization*. Some of the material was lemmatized automatically[6]. However, in order to

improve coverage, we performed a manual lemmatization[7]. 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 avoid the discussion of how fine-grained distinctions there is need for.

3. Preliminary results

There have been some different typologies of semantic change (for instance Bloomfield (1933), Ullmann (1963), Stern (1975), Blank and Koch (1999)). We will here follow the typology of Ullmann (1963)[8], giving examples of 'widening' and 'pejoration' of meaning.

3.1. *Widening*

One of the heuristics we have worked with is that words which are fashionable both as words but also as social concepts would have a difference in distribution. Examples of such words are 'relation', 'divorce' and 'social connection'. The word *kontakt* 'contact' is such an example in our data. Most striking is the difference in frequency of the lemma, with nine (sic!) occurrences in the 19th century data, and 1878 occurrences in the 20th century data.

We also see the difference in rankings, where the highest ranked verb-object construction with *kontakt* in 19th century data is on 7648th place (of 1870978 unique pairs), *koppla kontakt* 'connect contact'. Whereas the highest ranked verb-object construction with *kontakt* in the 20th century data is on 140th place (of 97484 unique pairs), *ta kontakt* 'take contact'.

The social sense of *kontakt* 'acquaintance' is a metaphorical transfer from the original source of contact between surfaces to the metaphorical contact between people. This is a widening. However, when it comes to metaphors usually *the whole picture* is transferred, not only single words. This leads to the result that even though there is a major increase in the number of occurrences, the verb-object constructions as such are not easily disambiguated in their minimal concordance form, many constructions can be of any sense. Hence, by looking at the verb distribution for *kontakt* it is difficult to establish a widening, thus we are left with the increased frequency as such, which heuristically is an indication of a semantic change.

3.2. *Emotive, pejorative change*

Surprisingly we do find a way to detect pejoration from a word that in many languages has gone from descriptive to pejorative. The word *neger* 'Negro' is actually slightly more frequent in the 19th century data (17 occurrences), even though there were fewer people of African descent in Sweden at that time. There are few occurrences in the 20th century data, only 15, and all verb-object constructions, except *säga neger* 'say Negro' (with four occurrences), has only one occurrence each. It seems to be mostly meta-linguistic, in that *neger* is occurring as the object of indirect speech verbs like *säga* 'say', *viska* 'whisper', *skrika* 'shout', *upprepa* 'repeat' and *diskutera* 'discuss', whereas for the 19th century data this is not a prominent feature (though occurring), and *neger* seem to be used descriptively.

4. Future work

The outcome of the present work can be a starting point for automatically detecting semantic change. We have here only briefly shown how differences in ranking and frequency lead us to suspect a change of sense. By comparing (fairly) comparable corpora from different time periods we can be made aware of changes, and might in some cases even track where the sense starts being polysemous and where the new sense possibly precedes the older sense in frequency. We will also compare lexical sets where the given words are within semantically close domains, which presumably will render further input in the pursuit of semantic change.

The manually made lemmatization is a valuable resource in enhancing the searches of the 19th century material in the Swedish Literature Bank, and can be used for building better automatic lemmatization on other Swedish material. By adding lemma and syntactic information to data, though labour intensive, we can make more reliable proposals concerning semantics.

Notes

[1] This follows the work of Jezek and Lenci (2007).
[2] In Swedish *läsa* is, besides other subsenses, ambiguous between 'study' and 'read'. In this case it is most certainly the sense of 'studying law'.
[3] There is ongoing work, independent of this work, to prepare the Swedish Literature Bank inclusion in Korp (Språkbanken, 2011).
[4] We have not yet made any proper error analysis.
[5] Personal communication Sofie Johansson Kokkinakis.
[6] This was carried out by Markus Forsberg, employing a similar approach to that used in Borin (2010).
[7] This work was partly funded by *Stiftelsen Erik Wellander*'s foundation.
[8] of *narrowing*, *widening*, *pejoration* and *amelioration*.

References

Banerjee, S. 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*, 370–381.

Blank, A. and P. Koch (eds.) 1999. *Historical semantics and cognition. Cognitive linguistics research 13*. Berlin: Mouton de Gruyter.

Bloomfield, L. 1933. *Language*.

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

Brants, T. 1998. 'TnT-Statistical Part-of-Speech Tagging.' http://www.coli.uni-sb.de/~thorsten/tnt/

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

Harris, Z. 1968. *Mathematical structures of language.* John Wiley & Sons.

Jezek, E. 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.

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

Michel, J.-B., 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.

Nivre, J. 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)*, 137–148.

Rohrdantz, C., 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. Stroudsburg, PA, USA: Association for Computational Linguistics, 305–310.

Sagi, E., 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: GEometical Models of Natural Language Semantics*. Athens, Greece, March 2009, 104–111

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

Sahlgren, M. 2008. ‘The distributional hypothesis.’ *Italian Journal of Linguistics* 20.1: 1–18.

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

Stern, G. 1975. *Meaning and change of meaning: with special reference to the English language*. Indiana University Publications Series. Greenwood Press.

Ullmann, S. 1963. *The principles of semantics*. Glasgow University publications. B. Blackwell.