Two cases of meaning change in Finnish newspapers, 1820-1910

Antti Kanner

Abstract for Nordic Digital Humanities Conference in Helsinki, March 2018

In Finland the 19th century saw the formation of number of state institutions that came to define the political life of the Grand Duchy and of the subsequent independent republic. Alongside legal, political, economic, and social institutions, Modern Finnish, as an institutionally standardised language, can be seen in this context. As the majority of residents of Finland were native speakers of Finnish dialects, adopting Finnish was necessary for state's purposes in extending its influence within the borders of the autonomous Grand Duchy. Widening domains of use of Finnish also played an important role in the development of Finnish national identity. In the last quarter of 19th century, Finnish started to gain ground as the language of administrative, legal, and political discourses alongside Swedish. It is during this period that we find most of the crucial conceptual processes that shaped Finnish political history.

In this paper I present two related case studies from my doctoral research, where I seek to understand the semantic similarity scores of so-called Semantic Vector Spaces in terms of linguistic semantics. The vector spaces have been obtained from large historical corpora of Finnish newspapers. Historical corpora are best understood as collections of past speech acts and the view they provide to changing meanings of words is shaped by contextual and pragmatic factors present at the moment of a texts' production. For this reason, understanding and explicating the historical context of observed processes is essential when studying temporal dynamics in semantic changes. To this end, I will try to reflect the theoretical side of my work in the light provided by actual cases of historical meaning changes. My research falls under the heading of Finnish Language, but is closely related to intellectual and conceptual history and computational linguistics.

The main data for my research comes from the National Library of Finland's Newspaper Collection, which I use via the KORP service API provided by Language Bank of Finland. The collection contains nearly all newspapers and periodicals published in Finland from 1771 to 1910, and Finnish publications from 1820. The collection is, however, very heterogenous, as the press and other forms of printed public discourse in Finnish only developed during the 19th century. Historical variation in conventions of typesetting, editing, and orthography, as well as paper quality used for printing make it difficult for OCR systems to recognize characters with 100 percent accuracy. Kettunen et. al. estimated that OCR accuracy is actually somewhere between 60 and 80 percent. However, not all problems in the automatic recognition of the data come from OCR problems or historical spelling variation. Much is also due to linguistic factors: the 19th century saw large scale dialectal, orthographical, and lexical variation in written Finnish. To exemplify the scale of variation, when a morphological analyser for Modern Finnish (OMORFI, Pirinen 2015) was used, it could only

parse around 60 percent of the wordlist of the Corpus of Early Modern Finnish (CEMF). For these reasons (unreliable results from automated parser and the temporal heterogeneity inherent in the data), conducting a methodology robust study poses a challenge. To mitigate these issues, the approach chosen was to use a number of analyses and see whether results could be combined to produce a coherent view of the historical change in word use.

All of the analyses in this work are based on term-feature matrices, of which term-document, term-collocation and term-morphological case category can be said to be of different types. Depending on specific tasks, methods in computational linguistics, such as LDA (Blei, Ng & Young 2004) or word2vec (Mikolov et. al. 2013), select one type of term-feature matrix as the starting point and then process this matrix into a more concentrated, embedded vector space. In research, most attention is usually reserved to the algorithm and selection of the type of original feature matrix in usually treated as more or less trivial matter. However, eg. Levy & Goldberg (2014) have noted that linguistic regularities observed in the embedded vector spaces are not consequences of the embedding process, but that the embedding process preserves those regularities well. Also, earlier research claims that the different ways the word representations are built from corpus data (ie. the specific type of the word-feature matrix used) seems to measure different types of semantic relatedness as semantic similarity (Sahlgren 2006, 2008). The methodological choices of this work, then, stem from these two observations. No embedding algorithms are used, and an array of analyses based on different types of word-feature matrices, is composed to monitor different semantic relations for semantic changes.

While a number of analyses based on different types of word-feature matrices were conducted (such as varying ngrams and skip-grams), two analyses deserve further discussion here. First, an analysis based on term-document matrices, such as topic modelling, seems to describe meaning relations that could be said to be associative, schematic, or discursive. In this study, this analysis was conducted using second order collocations (Bertels & Speelman 2014 and Heylen, Wielfaerts, Speelman, Geeraerts 2014) instead of algorithms like LDA, that are widely used for purpose of topic modelling. Preliminary tests showed that in this specific case, LDA was not able to produce well-formed topics in comparison to results from clustering second order collocations. This simpler approach seemed to be robust for some properties in the data that yielded LDA unusable, while it was left unclear what those properties actually were. Second, an analysis based on syntactic features was conducted by substituting syntactic dependencies with case marking distributions. This can be done on the grounds that the case selection in Finnish is mostly governed by syntax, as case selection is used to express syntactic relations between, for example, constituents of nominal phrases or predicate verb and its arguments (Vilkuna 1989). As automated syntactic parsing relies on preprocessed morphological analysis, fragility of syntactic parsing to errors in morphological preprocessing is of second order of magnitude. The aggregated morphological distributions on the other hand seem to be quite robust with regard to mistakes in the data, as the nature of noise the errors introduced is, in most cases, quite uniform. When the task is to track signals

of change, morphological case distributions can be used as sufficient proxies for dependency distributions.

The first of my case studies focuses on the Finnish word *maaseutu*. After its introduction to Finnish in the 1830's, *maaseutu* was used in more variety of related meanings, mostly referring to specific rural areas or communities. Staring from the 1870s it developed into a collective singular, referring to countryside as an undivisive whole and frequently contrasted to the urban, often lexicalised as *kaupunki*, the city. At the time when the the collective singular emerges, we find a number of occurrences which are vague in respect to specificity and collectivity.

Combining information from my analysis to newspaper metadata yields an image of a dynamic situation. The emergence of the collective singular stands out clearly, and is connected to an accompanying discourse of negotiating urban-rural relations on a national instead of regional level. This change can be pinpointed quite precisely to 1870s, and to newspapers with geographically wider circulation and a more national identity.

The second word of interest is *vaivainen*, an adjective referring to a person or a thing either being of wretched or inadequate quality, or suffering from a physical or mental ailment. When used as a noun, it refers to a person of very low and excluded social status and extreme poverty. The word has a biblical background, being used in older Finnish Bible translations (in, for example, the Sermon on the Mount as the equivalent of *poor* in Matt. 5:13: "blessed are the poor in spirit"), and as such was a natural choice to name the recipients of church charities. When the state poverty relief system started to take its form in the mid 19th century, it was built on top of earlier church organizations (Von Aerschot 1996), thus church terminology was carried over to these state institutions. Today, however, the word appears in Modern Finnish mostly in poetically archaic or historical contexts. It has disappeared from the vocabulary of social policy or social legislation by the early 20th century.

When tracking the contexts of the word over the 19th century using context word clusters based on second order collocations, two clear discoursal trends appear: the poverty relief discourse, that already in the 1860's is pronounced in the data, disperses into a complex network of different topics and discursive patterns. As state run poverty relief institutions become more complex and efficiently administered, the moral foundings of the whole enterprise are discussed alongside reports of everyday comings and goings of individual institutions or, indeed, tales of individual relief recipient's fortunes. The other trend involves the presence of religious or spiritual discourse which, against preliminary assumptions does not wane into the background, but experiences a strong surge in the 1870s and 1880s. This can be explained in part by the growth of revivalist Christian publications in the National Library Corpus, but also by the intrusion of Christian connotations into the political discussion on poverty relief systems. It is as if the word *vaivainen* functions as a kind of lightning rod of Christian morality in public poverty relief discourse.

While the methodological contributions of this paper are not highly ambitious in terms of language technology or computational algorithms, the combination of a complementary array of analyses instead of a methodology based on a single highly complex and opaque algorithm, might be seen to show an innovative approach to Digital Humanities. I argue that robustness and simplicity of methods makes the overall workflow more transparent, and this transparency makes it easier to interpret the results in wider historical or linguistic contexts. This allows us to ask questions which are not confined to the fields of computational linguistics or lexical semantics, but apply to wider areas of Humanities scholarship. This shared relevance of questions, intersections of interests of knowledge, to my understanding, lies at the core of Digital Humanities.

References

Bertels, A. & Speelman, D. (2014). "Clustering for semantic purposes. Exploration of semantic similarity in a technical corpus." *Terminology* 20:2, pp. 279–303. John Benjamins Publishing Company.

Blei, D., Ng, A. Y. & Jordan, M. I. (2003). "Latent Dirichlecht Allocation." *Journal of Machine Learning Research* 3 (4–5). Pp. 993–1022.

CEMF, *Corpus of Early Modern Finnish.* Centre for Languages in Finland. http://kaino.kotus.fi

Heylen, C., Peirsman Y., Geeraerts, D. & Speelman, D. (2008). "Modelling Word Similarity: An Evaluation of Automatic Synonymy Extraction Algorithms." *Proceedings of LREC 2008*.

Huhtala, H. (1971). *Suomen varhaispietistien ja rukoilevaisten sanankäytöstä : semanttis-aatehistoriallinen tutkimus*. [On the vocabulary of the early Finnish pietist and revivalist movements]. Suomen Teologinen Kirjallisuusseura.

Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T. & Kervinen, J. (2014). "Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods". In *IFLA World Library and Information Congress Proceedings : 80th IFLA General Conference and Assembly*. Lyon. France.

Levy, O. & Goldberg, Y. (2014): "Linguistic Regularities in Sparse and Explicit Word Representations." In *Proceedings of the Eighteenth Conference on Computational Language Learning.* Pp. 171-180.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. 2013: "Efficient Estimation of Word Representations in Vector Space." In arXiv preprint arXiv:1301.3781.

Pirinen, T. (2015). "Omorfi—Free and open source morphological lexical database for Finnish". In *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015*.

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

Sahlgren, M. (2008): "Distributional Semantics". *Rivista di Linguistica* 20:1. Pp. 33-53.

Vilkuna, M. (1989). *Free word order in Finnish: Its syntax and discourse functions*. Suomalaisen Kirjallisuuden Seura.

Von Aerschot, P. (1996). *Köyhät ja laki: toimeentukilainsäädännön kehittyminen kehitys oikeudellistusmisprosessien valossa.* [The poor and the law: development of Finnish welfare legislation in light juridification processes.] Suomalainen Lakimiesyhdistys.