Concepts Through Time: Tracing Concepts in Dutch Newspaper Discourse (1890-1990) using Word Embeddings

Melvin Wevers — Tom Kenter — Pim Huijnen

In this paper, we use a new technique, called Concepts Through Time (CTT), to trace concepts in newspaper discourse. CTT makes use of sequential semantic spaces to follow semantic shifts of concepts diachronically. The semantic spaces are based on the extensive digitized newspaper corpus of the Dutch National Library.[1] *As part of our approach, concepts are modeled as words related to each other in a semantic space. A key aspect of CTT is that the vocabulary used in the debate on a concept may change over time. In fact, the set of words used to discuss a particular concept might not show any overlap at all between different periods of time. As such, our method bears a resemblance to the rhizomatic structure as described by philosopher Gilles Deleuze. The conceptual implication of this method approximates the traditional scholarly methods of conceptual history. We test our method through three case studies: consumerism, globalization, and economic models. These domains are inextricably linked to the manifestation of modernization and*

[1] This collection includes over half a million newspapers for the period between 1890 and 1995. See www.delpher.nl

Americanization within Dutch public discourse throughout the twentieth century.[2] *Tracing the genealogy of these manifestations helps us to rethink the variety of meanings of modernization and Americanization in the Netherlands during this period. Essential to the development and success of this method is that, for every step of the process, decisions have been motivated from both the perspective of computational linguistics and cultural history. This makes this research project a genuine collaborative digital humanities effort.*[3]

Traditionally, the field of conceptual history has tried to combine micro and macro perspectives by studying the emergence and transformation of concepts, ideas, and thoughts over larger periods of time. Historians have increasingly used digital tools for the purposes of conceptual history. However, researchers in these studies regularly employed pre-defined and ahistorical definitions of concepts. Full-text search and n-gram viewers, for example, require a workable definition of a concept or a range of words that cover the subject in order to, subsequently, analyze them within certain contexts and periods.[4] The necessity of pre-defining terms is a serious drawback of working with these tracking tools. The research done in this way runs the risk of ahistoricity. The same goes for top-down approaches, in which a specific language model allows for the recognition of specific semantic information, such as entities via Named Entity Recognition[5] or via word classification

---

[2] This paper is part of the research project Translantis, which looks into the role of the United States as a reference culture in Dutch public discourse. Therefore, the chosen case studies are related to issues of Americanization within the field of consumer society and economy.

[3] Collaboration between computational experts and humanities scholars is an elemental part of the digital humanities. See: Anne Burdick et al., eds., *Digital Humanities* (Cambridge: MIT Press, 2012), 15–17.

[4] This method was adopted to study eugenics by Pim Huijnen et al., "A Digital Humanities Approach to the History of Science," in *Social Informatics*, ed. Akiyo Nadamoto et al., Lecture Notes in Computer Science 8359 (Springer Berlin Heidelberg, 2014), 71–85.

[5] Claire Grover et al., "Named Entity Recognition for Digitised Historical Texts.," in *LREC*, 2008, http://ltg.ed.ac.uk/np/publications/ltg/papers/bopcris-lrec.pdf; Seth van Hooland et al., "Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections," *Literary and Linguistic Computing*, November 29, 2013, http://llc.oxfordjournals.org/content/early/2013/11/29/llc.fqt067.

lists.[6] Topic modeling approaches partly circumvent this limitation by modeling groups of words rather than single words. Topic modeling tools such as Mallet and TMT allow users to find topics in a corpus of texts without having to provide the algorithm with any prior information.[7] With approaches such as "Topics over Time" researchers could also track topics diachronically.[8] Other research projects have used topic models to develop contextualized dictionaries that helped, e.g., to define the concept of 'strikes' in twentieth century Netherlands[9] or the Chicago School of 'neoliberalism' in postwar Germany.[10]

The methods described above all share the limitation that concepts – whether defined by hand or automatically – are static. Either the user provides a fixed list of words, or a fixed list of topics is inferred from the data in these topic-tracking approaches. However, historians commonly use texts to get a grip of the continuities and discontinuities throughout time. This involves the recalibration of their hermeneutical framework as they go along, i.e. the list of words they work with. We propose that digital tools should also be able to provide argumentative evidence for such a methodology in constructing historical narratives.

For this purpose, we introduce a new technique, called CTT (Concepts Through Time) that

---

[6] S. Klingenstein, T. Hitchcock, and S. DeDeo, "The Civilizing Process in London's Old Bailey," *Proceedings of the National Academy of Sciences* 111, no. 26 (July 1, 2014): 9419–24, doi:10.1073/pnas.1405984111.

[7] For an explanation of Topic Modeling see: David M. Blei and John D. Lafferty, "Topic Models," *Text Mining: Classification, Clustering, and Applications* 10 (2009): 71. For uses of topic modeling in historical research see: David J. Newman and Sharon Block, "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper," *Journal of the American Society for Information Science and Technology* 57, no. 6 (2006): 753–67; David Mimno, "Computational Historiography: Data Mining in a Century of Classics Journals," *J. Comput. Cult. Herit.* 5, no. 1 (April 2012): 3:1–3:19; Cameron Blevins, "Topic Modeling Martha Ballard's Diary," *Historying*, April 1, 2010, http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/; Peter Wittek and Walter Ravenek, "Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling.," 2011, http://bada.hb.se/handle/2320/9689.

[8] Xuerui Wang and Andrew McCallum, "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2006), 424–33, http://dl.acm.org.proxy.library.uu.nl/citation.cfm?id=1150450.

[9] Toine Bogers and Antal Van den Bosch, "Recommending Scientific Articles Using Citeulike," in *Proceedings of the 2008 ACM Conference on Recommender Systems* (ACM, 2008), 287–90, http://dl.acm.org/citation.cfm?id=1454053.

[10] Gregor Wiedemann, Andreas Niekler, and others, "Document Retrieval for Large Scale Content Analysis Using Contextualized Dictionaries," in *Terminology and Knowledge Engineering 2014*, 2014, http://hal.archives-ouvertes.fr/hal-01005879/.

aims to resolve the deficiencies of earlier methods. The technique is explicitly developed to account for historical changes in the semantics of concepts, thereby approximating the way historians traditionally have worked within conceptual history. CCT enables historians to trace concepts over large periods without having to manually select appropriate terms for the entire time span and without being dependent on a fixed set of topics. This allows for a greater sensitivity to semantic changes and an increased interactive heuristic approach to concepts within their discursive context.[11]

Methodologically, CTT is based on word embeddings, as inferred by a neural network trained on a large body of data, to monitor semantically related words. These are created using word2vec[12], a computational technique that does not depart from a top-down model of language. Rather, a semantic space is inferred from the input data. Word2vec uses a continuous bag-of-words or a skip-gram architecture in order to produce a multi-dimensional word-vector space. This space contains semantic and linguistic regularities that can be used for the analysis of discourse.[13] A positional shift within the vector space can been established as an indicator for chronological shifts on a semantic and syntactic level, for example in the use of the words “gay”, and “cell.”[14] This is not done using traditional means of analysis, such as part-of-speech tagging, but by merely observing the geometry in the vector space. However, the authors who have provided the mentioned

---

[11] The importance of context for historical research is eloquently expressed by Pelle Snickars, “If Content Is King, Context Is Its Crown,” *VIEW Journal of European Television History and Culture* 1, no. 1 (February 21, 2012): 34–39.

[12] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in Neural Information Processing Systems*. 2013.

[13] Marco Baroni Georgiana Dinu Germán Kruszewski, “Don’t Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors,” accessed September 11, 2014, http://anthology.aclweb.org/P/P14/P14-1023.xhtml; Derry Tanti Wijaya and Reyyan Yeniterzi, “Understanding Semantic Change of Words over Centuries,” in *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web* (ACM, 2011), 35–40, http://dl.acm.org/citation.cfm?id=2064475.

[14] Yoon Kim et al., “Temporal Analysis of Language through Neural Language Models,” *arXiv:1405.3515 [cs]*, May 14, 2014, http://arxiv.org/abs/1405.3515. In this paper, the authors show how gay has shifted from the connotation with an emotion to sexuality, cell has shifted from the denotation of a prison cell to the cellular phone.

examples have worked with single words of which the common meaning has shifted almost diametrically. We will be looking at clusters of words and shifts of meaning both large and small.

The data we use to trace concepts are the circa 500,000 newspaper issues between 1890 and 1990 that are available in the Dutch National Library's digitized newspaper archive. We train subsequent semantic model spaces in which we trace groups of terms, rather than individual words, diachronically and synchronically by keeping track of semantic relations between terms per period.

For this approach, no pre-established and fixed topic sets is needed. We will use a single seed set of terms, merely as an entry-point into the cluster to then find semantically related words within the semantic model. We will automatically update the set of related words over time utilizing the subsequent semantic spaces. A key aspect of this procedure is that the original seed words might disappear from the cluster of words over time.

The way our word embedding technique clusters words, holds surprising similarities with philosopher Gilles Deleuze's notion of the rhizomatic structure.[15] Important features of this structure are that it allows for multiple entry and exit points, and that the semantic meaning of a concept is determined through its "relations of exteriority."[16] Translated to our approach, the meaning of a term can be considered in terms of the geometric relation in a semantic space.[17] Starting from this theory, we aim to focus on the stabilization and destabilization of relational vectors between words, i.e. the emergence and disappearance of words within a subset, as well as

---

[15] Gilles Deleuze, *A Thousand Plateaus: Capitalism and Schizophrenia* (University of Minnesota Press, 1987), 6–23; For a clear description of rhizomatic thinking and assemblage theory see "Assemblage Theory," Texas Theory Wiki, *Assemblage Theory*, (2010), http://wikis.la.utexas.edu/theory/page/assemblage-theory. An interesting implementation of Deleuze's notion of assemblage and rhizome has been performed by Manuel DeLanda, *A New Philosophy of Society: Assemblage Theory and Social Complexity* (London: Continuum, 2006).

[16] DeLanda, *A New Philosophy of Society*, 10–11.

[17] Michel Foucault makes a similar point when he argues that the meaning of an expression cannot be simply deduced from syntactic and semantic meaning. Rather, one should look into the conditions in which expressions appear and change, in his words "by the analysis of the relations between the statement and the spaces of differentiation, in which the statement itself reveals the differences." See: Michel Foucault, *The Archaeology of Knowledge [1974]* (London: Random House, 2012), 105.

the shifting position of words in relation to one another. The latter could be the effect of forces within the cluster or changing vector relations outside of the cluster. We argue that the semantic space allows us to analyze "the processes that historically produce the identity of a given whole, but also the processes [coding and decoding] that maintain that identity through time"[18] Using this approach, we hope to determine words central to a topic and avoid concept drift caused by ambiguous words.

The development of this tool stems from a close collaboration between information retrieval experts and cultural historians. Every single technical or methodological choice has been motivated such that it complied with both sound computational linguistic research and critical historical methods. Consequently, we have high hopes of the applicability of this technique in historical research. We will evaluate our approach through three historical use cases in the context of the historiographical debates on modernization and Americanization in the Netherlands.[19] In the first case, we will trace popular consumer goods, such as for instance cigarettes, alcohol, and fast food. This case study sets out to show which consumer goods appeared in newspapers discourse in specific periods. Moreover, it might show us in what wider discursive contexts these consumer goods were discussed, e.g. as luxury goods or as unhealthy products. The second use case maps out businesses in newspaper discourse. This might shed light on processes of globalization in which local businesses are substituted by multinationals. The final use case deals with the notion of efficiency. The concept was introduced in the Netherlands after the First World War[20]; we will trace the genealogy of this concept, to see whether the concept was already present in earlier discourse

---

[18] DeLanda, *A New Philosphy of Society,* 10.

[19] These use cases stem from the Translantis research project. See www.translantis.nl and Joris van Eijnatten, Toine Pieters, and Jaap Verheul, "Big Data for Global History: The Transformative Promise of Digital Humanities," *BMGN - Low Countries Historical Review* 128, no. 4 (December 16, 2013): 55–77.

[20] Erik Bloemen, *Scientific Management in Nederland, 1900-1930* (Leiden: Doctoral thesis, 1988).

without being denoted as ‘efficiency.’ In our final paper, we will show a formalization of the search strategies used to trace concepts through time. Furthermore, we will reflect on the representativeness of the newspaper archive as well as the applicability of the method to other corpora.