Tracing Concepts Research Plan

# Introduction

This research project applies techniques developed in the ShiCo project to other datasets and aims to evaluate word embeddings techniques for historical research.[[1]](#footnote-0) The main outcome will be a white paper documenting these steps. ShiCo uses word embeddings to trace conceptual change over time. These word embeddings were trained with the word2vec algorithm using a shifting window approach on ten-year slices of Dutch newspaper data.[[2]](#footnote-1) ShiCo was trained using default settings on a rather messy dataset that included many OCR-errors as well as many different types of newspapers that consisted of both articles and advertisements. In this project, we aim to evaluate the training parameters and to determine to what extent the Shico algorithm can also be applied to different datasets to study conceptual change. As a result, we intend to define the best practices for selecting and preprocessing data, as well as for the training and evaluating of models. The evaluating consists of quantitative and qualitative assessment of the quality of the models. For the latter, we will use the input of domain experts.

# Conceptual History

Concepts acquire meaning from their use in specific historical contexts.[[3]](#footnote-2) The genealogy of a concept can inform us about social change and the ways we think about social change.[[4]](#footnote-3) The field that deals with semantic changes over time has also been named historical semantics.[[5]](#footnote-4) The research agenda deals with finding out who where the codifiers and shapers of meaning, and the discovery of breakpoints in the development of meaning. When did particular meanings became dominant, and what did others lose prominence? As Koselleck argues “coefficients of change and acceleration transform old fields of meaning, and, therefore, political and social experience as well.”[[6]](#footnote-5)

# Concepts Represented in Word Embeddings

Word embeddings are a type of word representation in which words with similar semantic and syntactic relationships have a similar representation in a multidimensional space. Word embedding algorithms, such as Word2vec, model relationships between words spatially based on co-occurrences of words within a set window span. In this multidimensional space, semantic and syntactic information is represented by geometry.[[7]](#footnote-6) This allows researchers to query the space using algebra, as represented by the famous example: King - Man + Women = Queen. Word embeddings offer a way to study words in their context—represented by nearest neighbors—in particular historical periods. By interlacing multiple models, we can even trace changes and stability in semantic context.

Concepts are actually represented as words or lists of vocabularies. Wittgenstein argues that the “meaning of a word is its use in the language.” By tracing the contextual shifts of words, we can “travel with the word’s uses through a complicated network of similarities overlapping and crisscrossing.”[[8]](#footnote-7) (Wittgenstein, PI 66). These networks of words reveal the “architecture of concepts” — “the words, phrases, sentences, and statements that we use and use us.”[[9]](#footnote-8) Shico offers a way to trace these networks and learn how certain associations emerged and evolved over time.

# Evaluating Word Embeddings

Recently, word embeddings models received ample attention, but scholars also expressed criticism for them being intangible.[[10]](#footnote-9) This project will apply several evaluation metrics, both intrinsic and extrinsic, on different datasets to better our understanding of how to train models for specific historical tasks, taking into account the specifics of a dataset. Moreover, we will discuss which types of questions can be answered using word embeddings models.

# Data

This project will use a number of different sources:

* Newspapers
  + Volkskrant
  + Parool
  + Trouw
  + Algemeen Handelsblad - NRC/Handelsblad
  + De Telegraaf
* Academic journals
  + Tijdschrift voor Geschiedenis

# Steps

* Prepare data
  + downloading data from repositories
  + decide on the pre-processing steps
    - remove case? (probably not)
    - lemmatization?
    - remove digits?
    - remove short words (shorter than 3 or 4 characters)
  + extract sentences from document and tokenize and divide into sentences
* Make models using word2vec
  + Window size -- how big will the slices be? Smaller windows allow for more fine-grained detection of change points, but more data results in better models.
  + Train sets using different settings (1 year, 2, and 5 year)
  + Different training parameters.[[11]](#footnote-10)
  + Translate analogy test google to Dutch
  + Determine seed words and determine use cases

# Evaluation

* Reliability: comparing the n next neighbors (by cosine distance) for each word modeled by the experiments with a variant of the Jaccard coefficient
* Accuracy
  + analogy set > calculate percentage of correct analogies and similarities
  + <https://github.com/mbatchkarov/repeval2016>
  + <http://veceval.com/>.
* Semantic relatedness reflects the degree to which two words share attributes (Turney et al. 2010, p. 149)
* Similarity is defined by Turney as co-hyponymy (car and bicycle), whereas hill et al. 2015 define it as the “words with identify referents (mug and cup)
* Task-based evaluation

# Budget and Plan

|  |  |  |  |
| --- | --- | --- | --- |
| **Activity** | **Responsible** | **Effort** | **Risks & solutions** |
| Data preparation | DHG/Huygens | .5PM | Encoding problems; solution: conversion to UTF-8.  Convert XML structure to raw text, solution: write parser. |
| Set up experimental pipeline | DHG | .5PM |  |
| Define evaluation criteria | DHG/Huygens | .5PM | Determine evaluation criteria and write scripts for them. Solution: collaboration with engineers Huygens and use input VU |
| Implement evaluation criteria in experimental pipeline | DHG | .25PM |  |
| Evaluate outcomes first round of experiments | Huygens/DHG | .25PM |  |
| Implement changes after first evaluation and run second experiments | DHG | .5PM |  |
| Evaluate outcomes second round of experiments | Huygens/DHG | .25PM |  |
| Write up report & paper | DHG/Huygens | 1PM |  |

**TO DO**

* Download and prepare data
* Gather examples of papers using word embeddings for historical research concept
* Train models
* Evaluate
* Write paper in Overleaf
* Check out code base: <https://github.com/williamleif/histwords>
* https://bitbucket.org/omerlevy/hyperwords/src/688addd64ca2ce8b4772be317f0b980f7716f4d6/scripts/?at=default
* Also check out alignment script in this github and Carlos’

[**https://gitlab.com/morlikowski/diachronic-analogies-code**](https://gitlab.com/morlikowski/diachronic-analogies-code)

**https://github.com/clips/dutchembeddings**

1. Kenter, Tom, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. “Ad Hoc Monitoring of Vocabulary Shifts over Time.” In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1191–1200. ACM, 2015. <http://dl.acm.org/citation.cfm?id=2806474>; Martinez-Ortiz, Carlos, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul, and Joris Van Eijnatten. “Design and Implementation of ShiCo: Visualising Shifting Concepts over Time.” In *HistoInformatics 2016*, 1632:11–19, 2016. [↑](#footnote-ref-0)
2. [Tomas Mikolov](https://arxiv.org/find/cs/1/au:+Mikolov_T/0/1/0/all/0/1), [Kai Chen](https://arxiv.org/find/cs/1/au:+Chen_K/0/1/0/all/0/1), [Greg Corrado](https://arxiv.org/find/cs/1/au:+Corrado_G/0/1/0/all/0/1), [Jeffrey Dean](https://arxiv.org/find/cs/1/au:+Dean_J/0/1/0/all/0/1) (2013) Mikolov et al., 2013Efficient Estimation of Word Representations in Vector Space. [**arXiv:1301.3781**](https://arxiv.org/abs/1301.3781) [↑](#footnote-ref-1)
3. Skinner, Quentin. *Visions of Politics*. Cambridge University Press, 2002; Koselleck, Reinhart, and Todd Samuel Presner. *The Practice of Conceptual History: Timing History, Spacing Concepts*. Stanford University Press, 2002. [↑](#footnote-ref-2)
4. James, Paul, and Manfred B. Steger. “A Genealogy of ‘Globalization’: The Career of a Concept.” *Globalizations* 11, no. 4 (July 4, 2014): 417–34. https://doi.org/10.1080/14747731.2014.951186. [↑](#footnote-ref-3)
5. Allan, Kathryn, and Justyna A. Robinson. *Current Methods in Historical Semantics*. Walter de Gruyter, 2012; Junge, Kay, and Kirill Postoutenko. *Asymmetrical Concepts after Reinhart Koselleck: Historical Semantics and Beyond*. transcript Verlag, 2014. [↑](#footnote-ref-4)
6. Koselleck, Reinhart, and Todd Samuel Presner. “On the Need for Theory in the Discipline of History.” In *The Practice of Conceptual History: Timing History, Spacing Concepts*, 1–19. Stanford University Press, 2002, 5. [↑](#footnote-ref-5)
7. Miller, George A., and Walter G. Charles. “Contextual Correlates of Semantic Similarity.” *Language and Cognitive Processes* 6, no. 1 (January 1, 1991): 1–28. <https://doi.org/10.1080/01690969108406936>; Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. “Don’t Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors.” In *ACL (1)*, 238–247, 2014. http://anthology.aclweb.org/P/P14/P14-1023.pdf. [↑](#footnote-ref-6)
8. Wittgenstein, Ludwig. *Philosophical Investigations*. John Wiley & Sons, 2010, 43 & 66. [↑](#footnote-ref-7)
9. De Bolla, Peter. *The Architecture of Concepts: The Historical Formation of Human Rights*. New York: Fordham University Press, 2013. [↑](#footnote-ref-8)
10. Hellrich, Johannes, and Udo Hahn. “Bad Company-Neighborhoods in Neural Embedding Spaces Considered Harmful.” In *COLING*, 2785–2796, 2016. http://www.aclweb.org/anthology/C/C16/C16-1262.pdf. [↑](#footnote-ref-9)
11. Hellrich, Johannes, and Udo Hahn. “Bad Company-Neighborhoods in Neural Embedding Spaces Considered Harmful.” In *COLING*, 2785–2796, 2016. <http://www.aclweb.org/anthology/C/C16/C16-1262.pdf>; https://sites.google.com/site/repevalacl16/ [↑](#footnote-ref-10)