Design and implementation of ShiCo: Visualising shifting concepts over time

Carlos Martinez-Ortiz[1], Tom Kenter[2], Melvin Wevers[3], Pim Huijnen[3], Jaap Verheul[3], and Joris van Eijnatten[3]

[1] Netherlands eScience Center, Amsterdam, the Netherlands
c.martinez@esciencecenter.nl
[2] University of Amsterdam, Amsterdam, the Netherlands
tom.kenter@uva.nl
[3] Utrecht University, Utrecht, the Netherlands
m.j.h.f.wevers@uu.nl, P.Huijnen@uu.nl,
J.Verheul@uu.nl, J.vanEijnatten@uu.nl

Abstract

In different times, people use different words to describe concepts. Change and stability in word usage are possible indicators of wider socio-cultural changes. To gain insight into how people perceive concepts, it is valuable to trace how the words denoting a certain concept change over time. Existing tools for exploring historical concepts, such as keyword searching or topic modeling, are ill-suited for the task; they are either too top-down or too rigid for an iterative exploration of historical concepts in large data sets.

In this article, we present ShiCo: a graphical interface for visualising concepts over time by monitoring shifts in word usage in a document corpus. As the dimension of time plays a crucial role in ShiCo, this article demonstrates ShiCo on a large corpus of newspaper articles spanning several decades. We describe the design choices made during the development of ShiCo and the key parameters that control the tool's behaviour. Lastly, as ShiCo is meant to be used by the broader community, we describe the steps required for running ShiCo on a novel data set.

1 Introduction

In different times, people use different words to describe concepts. Change and stability in word usage are possible indicators of wider socio-cultural changes. To gain insight into the ways in which people perceive concepts, it is valuable to trace precisely how words denoting a particular concept change over time. This paper introduces ShiCo, a graphical interface for visualising concepts over time by monitoring word usage in a a document corpus.

The study of concepts has a long tradition in the field of history. It reached maturity, particularly, within two schools: the Bielefeld school of *Begriffsgeschichte* associated with Reinhart Koselleck and the Cambridge School of Intellectual History connected to Quentin Skinner. For all their differences, the two schools have much in common. They both stress the historical situatedness of concepts as "the meanings of the words that convey them" [1]; both claim that the meaning of concepts depends on the historical contexts in which they occur, and on the contrast with contested meanings of similar words. For both schools, contextuality is what defines concepts in the first place - and sets them apart from words. In ShiCo, the notion of contextuality is incorporated by defining *semantic networks* of related terms [9, 2, 5]. Section 2.1 below presents further implementation details.

Computational techniques as simple as keyword searching are a valuable addition to the study of history if they are premised on the difference between words and concepts. Analysis

M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch (eds.): Proceedings of the 3rd HistoInformatics Workshop, Krakow, Poland, 11 July 2016, published at urlhttp://ceur-ws.org

of the (linguistic) contexts of concepts, such as race or progress in different periods is crucial to developing historical insight into both the continuity or stability of concepts, as well as the shifts in words and terms that people used to denote them.

To systematically measure the changes in words used to refer to concepts, we make use of a corpus of documents spanning an extended time period. In particular, we use a corpus of Dutch newspaper articles extracted from the Dutch newspaper archive[1] provided by the Dutch Royal Library (KB - Koninklijke Bibliotheek[2]) that were published between 1950 and 1990.

The remainder of this paper is structured as follows: Section 2 discusses the design decisions we made while building ShiCo, how these decisions influence its operation and how the ShiCo's output should be interpreted while taking the tool's strengths and weaknesses into account. Section 3 covers how other researchers can use ShiCo to analyze their own data sets. Section 4 presents examples that illustrate the functionality of ShiCo. Section 5 concludes with our final remarks and possibilities for future research.

2 Design and implementation of ShiCo

This section describes the design of ShiCo. We shall first describe the algorithm used to compute word semantics over time that underlies ShiCo. Then, we shall discuss the parameters that users can adjust in the graphic user interface.

2.1 Underlying algorithm for establishing word relatedness using semantic models

In [3] the algorithm underlying the ShiCo user interface is presented. A crucial notion in the algorithm is that of a *semantic model*, also referred to as a *semantic space*. In this semantic space, words are represented as vectors, and the distance between the vectors corresponds to their semantic similarity. These semantic models are derived from large amounts of unlabeled text. For ShiCo, multiple semantic models were calculated from newspaper articles published in multiple periods of time, using word2vec [6] (we use the Gensim implementation [8]). The time periods used span ten years and overlap to capture subtle changes in word usage.

To monitor changes in word usage over time, the algorithm uses two steps: generation and aggregation. The generation step employs an iterative approach. For an initial *seed set* (a small set of user-defined terms, typically 1 or 2 terms) the closest, i.e. semantically most similar, terms in the semantic space are retrieved. From these related terms, a semantic graph is constructed. Graph-based measures are used to find the most central terms in the semantic graph. The topmost central terms are used as the seed set for the next iteration of the algorithm. In the aggregation step, the lists of words produced in the generation step are aggregated over, to produce the final word lists, or *vocabulary*, which is then presented to the user.

Deciding which seed terms should be used depends on the research question the user wishes to investigate. The seed terms can be seen as the search terms in a search query. Useful seed terms could be selected based on the user's domain expertise or via automated algorithms [4].[3]

[1] The corpus can be queried at http://www.delpher.nl/

[2] See https://www.kb.nl/

[3] The authors thank the reviewers for pointing to this reference.

2.2 Application design

ShiCo consists of two main components: a RESTful backend and a web-based frontend. The backend implements the underlying *shifting vocabularies algorithm* described in the previous section; it holds the semantic models and generates the lists of terms related with a concept for each period. The frontend provides an interactive interface for generating shifting vocabularies and visualising them. The visualisation consists of two kinds of graphs: stream graphs and network graphs.

In the current implementation of ShiCo, the backend is written in Python while the frontend is a Javascript application using Angular. ShiCo is opensource software distributed under the Apache 2 Licence and publicly available via GitHub.[4]

2.3 Visualising time shifting vocabularies

As mentioned above, the frontend provides two kinds of graphs for visualising vocabularies over time. These graphs are complementary and provide insight into which terms are included in the vocabulary for each period, and above all, how terms in each epoch are related to each other. The graphs also indicate, for each period, the relative importance of terms (see Section 2.4 - Boost method). As such, the graphs provide users with a better understanding of why specific words are included in the results. Below, we shall provide a detailed explanation of the graphs:

- **Stream graph** – The stream graph (Figure 1a) shows a differently coloured stream for each term. These streams become broader or narrower depending on the relative importance of the term within each period.
- **Network graphs** – The network graphs (Figure 1b) display how terms within each period are related. ShiCo generates one network graph per period. Different symbols within the network graph indicate which words are seed terms (i.e. search terms for the semantic models), which are related terms (i.e. search results from the semantic models), and which terms were included in the final vocabulary (i.e. search results which have the highest relative importance in terms of graph-centrality).

2.4 Parameters

ShiCo's user interface provides multiple parameters that allow the user to control the behaviour of the underlying algorithm. This section describes the effect that these parameters have.

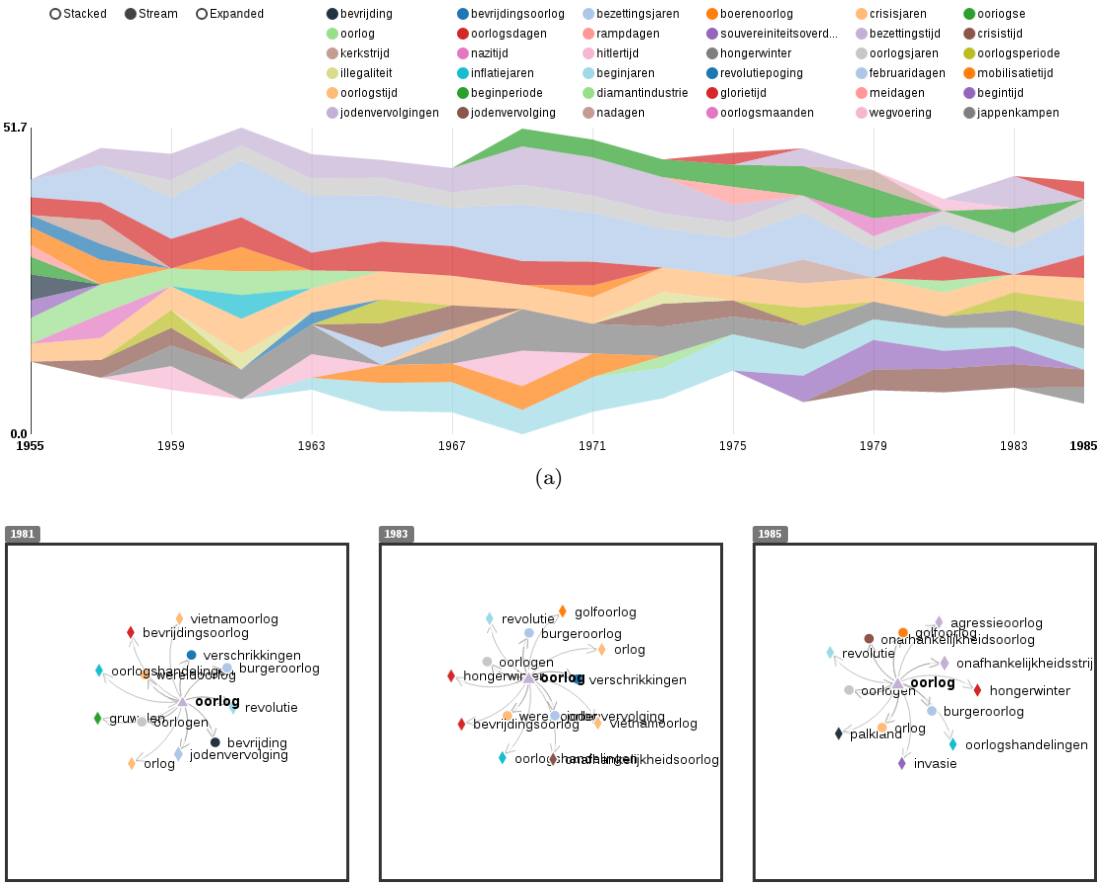
Boost method

The boost method is used to determine the relative importance of a term within its vocabulary. Two boost methods are available: (i) word counts, which takes into account the frequency of a word within each period (this is the relative frequency of a word within the vocabulary for each time period); (ii) sum of similarities, which takes into account the similarity of the word to the seed terms; smaller semantic distances indicate higher similarity between words.

Adaptive / nonadaptive tracking

Concept tracking can be performed in two modes: adaptive and nonadaptive. In adaptive mode, the terms resulting from searching in one semantic model become the seed terms for the

[4] https://github.com/NLeSC/ShiCo



(a)

(b)

Figure 1: Screenshot of ShiCo graphs

next semantic model (in chronological order) as described in Section 2.1. In nonadaptive mode, no iterative algorithm for updating the seed lists over time is applied. Instead, the same seed terms are used for all time periods.

Track direction

A concept can be tracked forward or backward through time. Searching forward means that a concept is followed as time progresses (for instance, we start with *Walkman* in 1980 and the model finds *discman* and *iPod* in later times). The process can also be reversed. That is, we might start with *iPod* and the underlying algorithm tracks this concept backward in time, to find *discman*, and earlier *Walkman*. Notice that this is only relevant for searches executed in adaptive mode.

Maximum concept distance

This refers to the semantic distance that must exist between a term and its seed word for the term to be included in the vocabulary. Words above the maximum semantic distance, which are

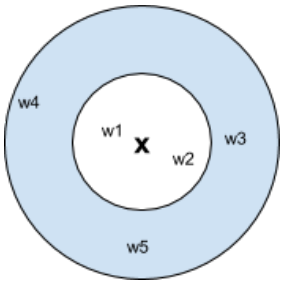


Figure 2: Illustration of maximum concept distance.

not sufficiently similar in the semantic space, are removed from the resulting vocabulary. This step is illustrated in Figure 2. For instance, a given seed term x produces a number of related words, $w_1 - w_5$. However, only the words which are within the maximum concept distance, w_1 and w_2, will be included in the final vocabulary.

Number of years in aggregation interval

Vocabularies resulting from individual models are aggregated to produce a smooth vocabulary that covers a longer time period. This aggregation reduces the inclusion of artifacts, which are words generated by a single model that do not occur in others. Users can define the number of years for these aggregation intervals. Using a higher number will reduce the noise in the results, at the risk of overlooking interesting but relatively weak events in the data. The optimum number of years in the data depends on the specific question being studied; researchers have the freedom to experiment with different values and draw their own conclusions from the results.

Maximum number of words per time period

ShiCo produces a vocabulary of related words for each period. Users can set a maximum number of words within each vocabulary. A smaller number of words will produce more focused results while a higher number will produce a broader concept search.

2.5 Performance optimization

ShiCo has been designed to perform tasks in parallel. Currently, in order to compute the lists of related words needed by the underlying algorithm described in Section 2.1, the most time-consuming task is large matrix multiplications – a task often encountered in computationally intensive research. Available software libraries, such as OpenBLAS [7], provide means to perform such multiplications in a parallelized way. ShiCo's performance greatly benefits from the use of such libraries, which reduce the time taken to perform a concept search.

Figure 3 illustrates how OpenBLAS improved the performance of ShiCo. This graph shows the distribution of processing time needed to complete each concept search from a set of 200 concepts. Each violin plot corresponds to a different configuration, using a different number of CPU's: C1 shows the performance using a single CPU, while C2, C4, C8, C16 and C32 show the distribution of times using OpenBLAS with 2, 4, 8, 16, and 32 CPU's, respectively. Each violin plot also shows (in red) the minimum, maximum, and average time of the concept search. There is large variability in the time taken to complete a search; this is most noticeable with

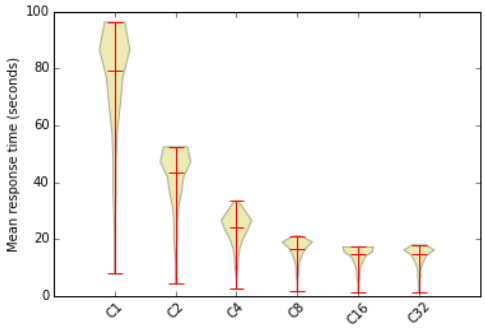


Figure 3: Performance improvement with OpenBLAS using multiple cores.

C1, where the slowest search takes close to two minutes to complete while the fastest search is completed in less than 10 seconds. The key feature of these plots is the mean time taken to complete a search. It can be observed that as the number of available CPU's increases, the search time decreases as expected. It can also be observed that, at a certain point, the gain of having multiple CPU's is overshadowed by the overhead of coordinating multiple threads: this is noticeable by the marginal gain between C16 and C32. At this point, doubling the number of CPU's yields a marginal improvement in performance.

It may be possible to apply further optimization techniques in order to improve ShiCo's performance. Additionally, the number of CPU's which achieve optimum performance may vary depending on the size of the semantic models that ShiCo is dealing with. Researchers are invited to explore and discover which would be the optimal settings for their data.

3 Use ShiCo on your data

ShiCo is open-source software. We encourage researchers to try it and use it as part of their research practice. This section provides an overview of the general requirements for using ShiCo on any dataset. Detailed instructions and technical information can be found on ShiCo's GitHub repository.

3.1 Data requirements

In order to use ShiCo on a given dataset, researches need to pre-process the data and compute semantic models from it. Ideally, sufficient data will be available to be able to create models spanning ten year periods. Unfortunately, it is impossible to specify exactly how many documents are needed to produce insightful results. However, generally speaking, more data will lead to more reasonable results.[5]

We prescribe grouping the data for the semantic models in ten year periods. The reason for this is twofold: (1) using ten year periods ensures that there is enough data for the semantic model; (2) ten year periods are sufficiently long to detect changes in word usage.

In cases when smaller (or larger) data sets are available, it may be possible to waive this requirement and aggregate the data in longer (or shorter) time periods. This will, of course, introduce the risk of overlooking a change in the vocabulary.

[5] As a reference, the corpus underlying the current ShiCo implementation contains over 26 million documents.

3.2 Data quality

In creating the semantic models, the quality of the data used is crucial. Data defects such as typographical errors and OCR errors will deteriorate the quality of the semantic models. For example, words may occur such as *ooriog* (which is meaningless in Dutch): a misspelling of the word *oorlog* (*war* in Dutch) caused by an OCR error.

3.3 Deploy front and backend

ShiCo is designed to be reusable regardless of a researcher's field of expertise. Starting up ShiCo is straightforward: the Python backend needs to be executed and the Javascript frontend needs to be provided with the URL indicating where the backend is running. As mentioned previously, ShiCo benefits from the availability of more CPU's, but in fact, it can also run (albeit somewhat slower) on a single-core machine. As to memory requirements, ShiCo requires enough memory to load all the semantic models into memory. Therefore, the total amount of memory required is linearly proportional to the number and size of the semantic models produced[6] – this may be another incentive to experiment with the parameters of the models.

3.4 Integration with other tools

As mentioned above, the newspaper dataset provided by the KB is used to power a currently operational instance of ShiCo. This instance is linked to another tool, called Texcavator,[7] which uses the same dataset. Texcavator is a text mining tool which enables historians to do full-text searches in document collections. Several visualisation options are available in Texcavator such as word clouds, time lines and metadata overviews. ShiCo has been integrated into Texcavator's user interface, and the two tools can be used together to explore the underlying data. This combination serves as an example of how ShiCo can be easily integrated with other tools.

4 Experimental results

This section describes a set of test cases which showcase the possibilities for data exploration using ShiCo.

4.1 Doping: forward/backward search

ShiCo offers the option for searching concepts forwards or backwards in time (in adaptive mode). In this example, we use the keyword *doping* (the English word is commonly used as a loan word in Dutch) as our seed concept, and we search forwards and backwards in time.

The forward search yields concept terms such as *hartaandoening* (heart condition), *bloedziekte* (blood disease), *longaandoening* (lung condition) and *nieraandoening* (kidney condition). These occurrences seem to indicate that the term doping was used to refer to drugs in the medical sense of the word.

The backward search yields concept terms such as *amphetaminen* (amphetamines), *anabole* (anabolic), *spierversterkende* (muscle-enhancing) and *amfetaminen* (amphetamines) after the 70's, shifting to terms such as *amfetaminen* (amphetamines), *morfine* (morphine), *cocaine* (cocaine) and *hasjiesj* (hashish) prior to that decade. These occurrences seem to indicate that our

[6] As a reference, the semantic models for the KB dataset currently use 32GB on disk, and about 50GB when loaded in memory.

[7] See: http://texcavator.surfsaralabs.nl/ – public access will be available in the future for both tools.

backwards search focuses on the use of illegal substances in sports and drifts towards illegal substances in general. The contrast between the forward and backward method suggests a possible shift in context in which doping was used in newspapers. The associated terms and their periodization in the vocabularies can be used to further trace the context of doping using other analytical tools.

4.2 Propaganda: Concept distances

ShiCo offers the option to control the maximum concept distance for a concept to be considered related to a seed concept. Modifying the maximum concept distance has an effect on the words which are related to a given search term.

In this example, we use the keyword propaganda as our seed concept, using different maximum concept distances.

Using a low maximum concept distance (0.3) ShiCo yields terms such as *protestdemonstraties* (protests), *haatcampagne* (hate campaign) and *prikacties* (pickets). These occurrences indicate that the search focuses on the political sense of the word.

Using a higher maximum concept distance (0.6) yields concept terms such as *reclame* (advertising), *commercieel* (commercial) and *redactie* (editorial). These terms indicate that the search focuses on the advertisement sense of the word.

Moreover, by analyzing the stream graph, we see that the initial associations of propaganda occurred predominantly just after the Second World War. The connotation to advertising becomes the dominant context of propaganda. In this case, we have monitored the conceptual drift, as the word propaganda moved from the context of politics.

5 Conclusion

In this paper we have presented a tool for visualising shifts in word usage over time. Examples from the KB dataset are used to illustrate how the tool works and how parameters in the user interface can be used to control the results obtained. ShiCo can be deployed with other data sets, and researchers are encouraged to do so.

ShiCo is an innovative tool which allows researchers to measure word usage changes over time and gain insight into changes in the concepts associated with such words. Researchers now have the opportunity to explore vocabularies related to concepts in an analytic and reproducible fashion. We hope this will lead to more robust and comparable discoveries in conceptual history and other areas.

There are of course limitations to what it is possible with ShiCo: as with any data driven approach, the quality of the data is directly related to the quality of the results. Researchers may also face a learning curve when initially using the tool, and interpreting results. Researchers are encouraged to develop an understanding of how the algorithms underlying ShiCo work in order to maximize the benefits from the tool.

Thus far, ShiCo has been tried with the KB dataset. Other data sets, in different languages and covering different time spans remain to be tested. Stream and network graphs have proven to be useful visualisations. However, other types of graphs might also provide valuable insight into the relations between words in vocabularies.

References

[1] P. de Bolla. The architecture of concepts: The historical formation of human rights. *Fordham University Press*, 2013.

[2] T. Kenter and M. de Rijke. Short text similarity with word embeddings. *In Proceedings of the 24th ACM international Conference on Information and Knowledge Management (CIKM'15)*, 2015.

[3] T. Kenter, M. Wevers, P. Huijnen, and M. de Rijke. Ad hoc monitoring of vocabulary shifts over time. *In Proceedings of the 24th ACM international Conference on Information and Knowledge Management (CIKM'15)*, 2015.

[4] Y. Kim, Y. Chiu, K. Hanaki, D. Hegde, and S. Petrov. Temporal analysis of language through neural language models. *CoRR*, abs/1405.3515, 2014.

[5] C. Martinez-Ortiz, T. Kenter, M. Wevers, P. Huijnen, J. Verheul, and J. van Eijnatten. Shico: A visualization tool for shifting concepts through time [demo]. *Digital Humanities Benelux*, 2016.

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013.

[7] W. Qian, Z. Xianyi, Z. Yunquan, and Q. Yi. Augem: Automatically generate highperformance dense linear algebra kernels on x86 cpus. *In the International Conference for HighPerformance Computing, Networking, Storage and Analysis (SC'13)*, 2013.

[8] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[9] M. Wevers, T. Kenter, and P. Huijnen. Concepts through time: Tracing concepts indutch newspaper discourse (1890-1990) using word embeddings. *In Digital Humanities 2015*, 2015.