Conceptual Change and Distributional Semantic Models: an Exploratory Study on Pitfalls and Possibilities

Pia Sommerauer and Antske Fokkens
Computational Lexicology and Terminology Lab
Vrije Universiteit Amsterdam
De Boelelaan 1105 Amsterdam, The Netherlands
pia.sommerauer@vu.nl, antske.fokkens@vu.nl

Abstract

Studying conceptual change using embedding models has become increasingly popular in the Digital Humanities community, while critical observations about them have received less attention. This paper investigates what the impact of known pitfalls can be on the conclusions drawn in a digital humanities study through the use case of "Racism" in the 20th century. In addition, we suggest an approach for modeling a complex concept in terms of words and relations representative of the conceptual system. Our results show that different models created from the same data yield different results, but also indicate that (i) using different model architectures, (ii) comparing different corpora and (iii) comparing to control words and relations can help to identify which results are solid and which may be due to artefacts. We propose guidelines to conduct similar studies, but also note that more work is needed to fully understand how we can distinguish artefacts from actual conceptual changes.

1 Introduction

Distributional models have been used to detect shifts in meaning with various degrees of success (Hamilton et al., 2016; Kim et al., 2014; Kulkarni et al., 2015; Gulordava and Baroni, 2011, e.g.). Based on promising examples such as the shift of the word *gay* from meaning 'carefree' to 'homosexual', researchers in digital humanities have been inspired to explore the use of distributional semantic models for studying the more complex phenomenon of concept drift (Wohlgenannt et al., 2019; Orlikowski et al., 2018; Kenter et al., 2015; Kutuzov et al., 2016; Martinez-Ortiz et al., 2016, e.g.). In most cases, standard methods with high results on identifying known examples of semantic shift are adopted and applied to specific data and use-cases.

Literature that raises critical questions concerning the reliability of these methods (e.g. (Hellrich and Hahn, 2016a; Dubossarsky et al., 2017)), however, seems to have received less attention in the digital humanities community. It is, in fact, far from trivial to apply distributional semantic models to study a complex phenomenon such as concept drift in a methodologically sound manner. We distinguish three main challenges: First, distributional semantic models reflect the way words are used and not directly how concepts are perceived. This leads to the question of which words should be studied and how shifts in their meaning relate to the underlying concept. Second, the relation between data, frequency and information emphasized by different model types is not fully understood (Dubossarsky et al., 2017). Third, the semantic models resulting from neural network-inspired architectures as provided by (e.g.) word2vec (Mikolov et al., 2013) depend on random factors such as initialization and the order in which data is presented (Hellrich and Hahn, 2016a).

If these challenges are not taken into account, researchers may end up publishing insights and results that are the result of artefacts in the data or models rather than valid observations on change. Existing research has shown that these variations exist, but we are not aware of previous work that explored their consequences in a typical digital humanities set-up, which does not just consider the most extreme changes or words in commonly used evaluation sets, but considers words of a specific topic under consideration. In order to enable digital humanities research that makes use of distributional semantic models, it is essential to establish how these models can be used in a methodologically sound manner and to communicate this to potential users.

In this paper, we illustrate this importance and

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 223–233
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

propose methods that take these risks into account when investigating conceptual change using word embeddings. We illustrate this through a use case of a concept known to have changed radically during the 20th century, namely "Racism". We define a set of words that represent this complex conceptual system and test various hypotheses concerning how relations between these words changed. We investigate the impact of artefacts by (1) using two datasets, (2) testing the impact on control words and (3) creating different models. In particular, we compare predict models both to count models and to other predict models created with different random initializations.

The results show that not all conclusions drawn in a naive methodological set-up can withstand a more critical investigation. The main contributions of this work are the following:

- We propsose ways of critically investigating apparent changes with respect to artefacts of the data and/or model.
- We formulate recommendations for Digital Humanities studies that aim to use diachronic embeddings to study conceptual change.

In addition, this paper provides a first illustration of how a generic hypothesis around a changing concept may be translated into concrete hypotheses concerning changes of language use.

We present this work in a somewhat unusual way to highlight the danger of uninformed use of distributional semantic models for studying concept drift. After an overview of related work (Section 2) and introducing our hypotheses (Section 3), we first take a naive approach using existing embeddings created according to the state-of-the-art and test our hypotheses in Section 4. We then report additional experiments that verify the robustness of the naively obtained insights in Section 5. Section 6 provides a set of recommendations on how to increase the reliability of research using distributional models to study language change based on the outcome and previous work. We then conclude and discuss open challenges.

2 Background and Related Work

Based on the distributional hypothesis (Firth, 1957), studying meaning change using distributional representations of words seems natural: Since words with similar meanings appear in similar contexts, it follows that changes in the contexts of words are a good indication of meaning change. This notion has been taken up in the Computational Linguistics community and implemented using distributional semantic models. The idea underlying diachronic distributional models is to create a series of semantic spaces representative of specific time periods that can be compared. While earlier approaches relied on count-based semantic space models (Gulordava and Baroni, 2011), more recent approaches made use of prediction-based models and suggested different methods to make embedding represention comparable across time periods (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016). Nowadays, prediction-based models (the skip-gram and CBOW architectures in the word2vec toolkit (Mikolov et al., 2013) and Glove (Pennington et al., 2014) seem to be the dominant choices (Kutuzov et al., 2018).

A number of studies warn about the reliability of distributional semantic models for detecting change. Dubossarsky et al. (2017) illustrate that it is not known what properties in the underlying corpora are emphasized by various models and that count-based models in particular are sensitive to frequency effects. Hellrich and Hahn (2016a) point out that predictive models trained on the same data return different nearest neighbors, because they are influenced by random factors such as their initialization and the order in which examples are processed. Antoniak and Mimno (2018) present an investigation of the extent to which only small changes in the underlying corpus impact the resulting representations. They show that the impact of the processing order increases when smaller corpora are used.

Researchers in other domains (mainly Digital Humanities, but also biomedical text minig (Yan and Zhu, 2018)) have embraced the promising initial results from studies such as Mitra et al. (2014) and Hamilton et al. (2016) without being aware of the pitfalls of these methods. This is particularly concerning, as these fields typically work with comparatively small datasets restricted to a specific domain (Wohlgenannt et al., 2019). For instance, Kenter et al. (2015) and Martinez-Ortiz et al. (2016) study conceptual change in a corpus of Dutch Newspapers collected by the Royal Dutch Libary. The same corpus is taken up by Orlikowski et al. (2018), who proposes a model of conceptual change using analogical relations between words. Kutuzov et al. (2016) extend the idea of diachronic changes to genre differences and explore subgenres of the BNC. Wohlgenannt et al. (2019) recognize problem of small specialized datasets and propose a new evaluation set constricted of data from the Game of Thrones and Harry Potter novels, but they do not address the problems related to robustness and frequency effects in their experimental set-up.

Even diachronic general purpose corpora, such as the Corpus of Historical American English (Davies, 2002, COHA) introduced to the Computational Linguistics community by Eger and Mehler (2016), are rather limited in size. The much larger Google n-grams data set (used by Mitra et al. (2014); Gulordava and Baroni (2011), e.g.) does not have this limitation, but full texts cannot be accessed and it suffers from a bias towards scientific publications from 1950 onwards (Pechenick et al., 2015). The Google n-grams fiction component, used by e.g. Michel et al. (2011); Dubossarsky et al. (2015), is smaller and limited in genre but avoids unbalanced differences in genre across time periods.

In addition to these model-specific caveats, the translation from (potentially complex) concepts to words which can be observed by a distributional model is not straight forward. Betti and van den Berg (2014) propose the use of conceptual models to study concept change in a clearly defined and somewhat formalized way. This notion is rarely treated explicitly in applications of diachronic embedding models. Studies such as Bjerva and Praet (2015) provide a start, but we are not aware of previous work that investigates a conceptual system consisting of several subconcepts and of a similar complexity to the use-case of "Racism" presented in this paper.

To the best of our knowledge, this work is furthermore the first to investigate how arteficial components influence a digital humanities research question. The scope of this research is still limited to investigating the impact of different methods and random artefacts leaving questions concerning the underlying data to future work.

3 Use Case: The Concept of Racism

The first step for studying concept drift by means of linguistic corpora is to identify words that refer to (components of) the concept and related concepts. Following Betti and van den Berg (2014)'s observation that change applies to conceptual networks, this can not be simplified by looking at words referring to the concept and their near synonyms alone. We distinguish four classes of words that can be relevant for studying conceptual change: (i) words referring to the core of the concept, (ii) relevant subconcepts, (iii) instances of a core or subconcepts and (iv) words referring to related concepts. In this paper we investigate how the concept of "Racism" changed during the 20th century. We use literature from various disciplines within Social Science and Humanities to select relevant words and formulate hypotheses. A brief overview is provided in this section.

Barker (1981) identifies a shift from 'old' to 'new' racism. Race used to be understood in biological terms related to visual attributes, particularly, skin color. Due to social changes (triggered by the Nazi regimes cruelties and the Civil Rights Movement), biological interpretations were relinquished as explanations for prejudice and increasingly replaced by cultural interpretations of differences between groups (Augoustinos and Every, 2007; Lentin, 2005; Morning, 2009; Omi, 2001; Wikan, 1999; Winant, 1998). We therefore identify "Culture" and "Race" as the core concepts of "Racism" investigated through the words *race* and *culture* as well as *racial* and *cultural* which are less polysemous. This shifting interpretation led to different ways of defining and comparing social groups (subconcepts and instances) and different justifications for racist ideologies (related concepts) summarized in Tables 1 and 2.

We hypothesize that words representing subconcepts, instances and related concepts associated with old racism will have moved further away (i.e. the similarity of their vectors has decreased) from the core concepts as this vision is no longer supported whereas words related to new racism have moved closer to the core concepts (i.e. the similarity between the vectors has increased) during the 20th century. Furthermore, we expect that within the core concepts, the word *cultural* is increasingly used to describe social groups, while the biologically connotated word *racial* is avoided. A detailed overview of all word pairs and their expected change can be found in the appendix to the paper (Appendix A).[1].

[1] Conceptual change in different corpora and models: https://github.com/cltl/semantic_space_navigation/tree/master/projects/conceptual_change, comparing models of the same corpus with different initializations: https://github.

Conceptual system of old racism		target words
Subconcepts	'Race defined in terms of visual attributes, first and foremost skin color	*skin color* (not investigated as compound nouns are not in the model vocabularies)
instances	Groups defined in terms of skin color	*whites*, *blacks*
Related concepts	Emphasis on a racial hierarchy	*superior*, *inferior*
	Biological justification of hierarchical structures	*genetics*
	Fear of intimacy between people of different racial groups	*marriage*, *relationship*

Table 1: Conceptual system and representative words of old racism.

Conceptual system of new racism		target words
Subconcepts	'Race' defined in terms of cultural background consisting of nationality, language and religion	*linguistic*, *national*, *religious*
instances	Group labels of immigrants	*immigrants*, *foreigners*
	Ethnic group labels	*Jews*, *Turks*, *Arabs*
Related concepts	Emphasis on differences	*different*
	Defense of seemingly liberal values	*values*, *attitudes*, *beliefs*
	The reason for differences lies in history (rather than genetics)	*historic*

Table 2: Conceptual system and representative words of new racism.

change direction	**ngrams**	**both**	**coha**
←→	*inferior - cultural* *superior- cultural*	*whites - races* *marriage - cultural*	
→←	*linguistic - cultural*	*values - cultures*	*religious - racial* *religious - racial* *different- cultural* *national - cultural*

Table 3: Hypotheses about changes in relations between words confirmed in the n-grams corpus, the COHA or both. The changes significantly correlate with time (either over the entire century or over the second half only).

4 Basic Experimental Results

In this section, we outline the outcome of a 'naive' approach to testing our hypotheses, using the methods with best results in Hamilton et al. (2016). We use two corpora: COHA with the advantage of being well-balanced and disadvantage of being relatively small (on average 24,5 million words per decade) and the larger but unbalanced English Google Ngram corpus (hencforth ngram). Embeddings are created by Hamilton et al. (2016) with the skip-gram with negative sampling model (SGNS) of the Word2vec toolkit.[2] We first explore whether cosine distances between vectors changed according to our hypotheses. Because we are ultimately interested in the reliability of positive results, we limit our presentation to the statistically significant confirmations presented in Table 3. We observe three hypotheses confirmed in both corpora, four only in the COHA corpus and three just in the Google Ngram corpus.[3]

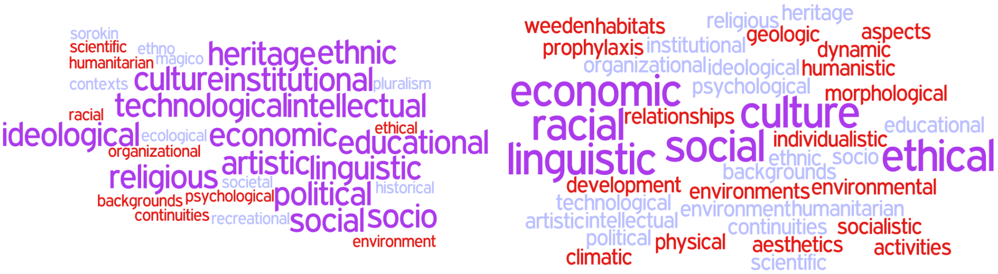
We furthermore explore changes in nearest neighbors of *cultural* and *racial* illustrated in Figures 1 and 2. The shifts observed in nearest neighbors indicate that biologically connotated term *racial* is increasingly avoided in contexts in which racially constructed groups are described or compared. The results indicate that it is used to name social problems partly rooted in racist ideologies.

This naive approach seems to confirm that the shift in "Racism" established by scholars is indeed reflected in language use to a certain extent. We observed stastically significant shifts between ten word pairs in the direction that was expected.

com/cltl/semantic_space_navigation/tree/master/projects/neighbor_stability

[2] The embeddings can be donloaded from the HistWords project webiste: https://nlp.stanford.edu/projects/histwords/

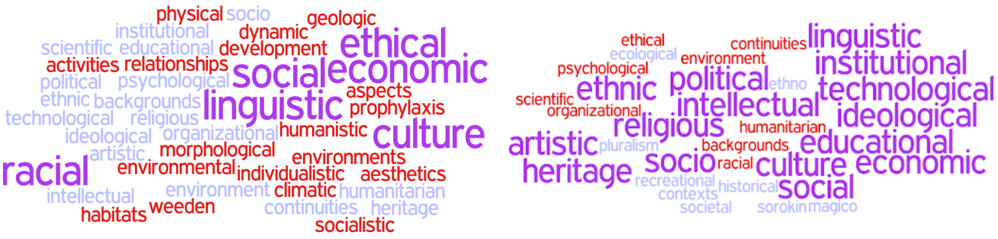
[3] Out of 47 hypotheses in total (see Appendix). A complete overview of the negative results is not included here due space limitations, but can be found in the Appendix.



(a) Nearest neighbors of *racial* in 1900 (small darker), 1950 (small lighter) and both decades (big).

(b) Nearest neighbors of *racial* in the 1950s (small darker), 1990 (small lighter) and both decades (big).

Figure 1: Changes in the nearest neighbors of *racial*.



(a) Nearest neighbors of *cultural* in 1900 (small darker), 1950 (small lighter), and both decades (big).

(b) Nearest neighbors of *cultural* in the 1950s (small darker), 1990 (small lighter) and in both decades (big).

Figure 2: Changes in the nearest neighbors of *cultural*.

We furthermore found changes in the environment of the nearest neigbors of *racial* and *cultural* that confirm the change of discourse from a biological racial vision of difference between people to a more cultural one. In the next section, we test whether the conclusions hold when being tested through alternative means.

5 Diving Deeper

At first sight, the approach and outcome outlined in the previous section may seem solid: we have taken the models evaluated best by Hamilton et al. (2016), who reported 100% accuracy on 18 evaluation pairs for the SGNS models created on the Google corpus. However, these results do not take into account that (1) predictive models are influenced by random components as pointed out by Hellrich and Hahn (2016a) and (2) significant change can also be spotted for words that did not exhibit change as (a) observed in the top-10 changing words reported in Hamilton et al. (2016) and (b) by the Dubossarsky et al.'s 2017 experiments showing that change is difficult to distinguish from frequency effects.

In this section, we present the results of additional experiments to test whether our initial findings hold when tested with alternative models. In addition, we use control words to verify whether changes between words referring to instances of racial groups and core concepts reflect indeed a change between these instances and the concept or whether similar changes are observed between the concepts and unrelated words or pairs of words whose distance should have remained stable.

5.1 Variations between Models

We first test whether a subset of our conclusions hold as well when we use Hamilton et al.'s 2016 count-based distributional semantic models, which are provided with their paper: a PPMI (Positive Pointwise Mutual Information) model and its high-density derivative SVD (Singular Value Decomposition). Though these models were less successful in detecting change in Hamilton et al.'s 2016 paper, they reflect the data directly without being influenced by their initialization or the or-

word pair	**SGNS**	**PPMI-SVD**	**PPMI**
culture-values	→←	→←	→←
races-immigrants	←→	←→	–
cultural-different	–	–	←→
racial-different	–	←→	←→
cultural-inferior	←→	←→	→←

Table 4: Comparison across different models using the ngrams corpus.

der in which examples are processed (Hellrich and Hahn, 2016b). Table 4 presents an overview of the conclusions drawn from different model types when analyzing changes in the relations between word pairs. Some changes are only significant in one model (e.g. *cultural-different*), others reveal contradictory results with significant changes in opposite directions (e.g. *cultural-inferior*). A conclusion that remains stable and is thus supported by all models is the increasing similarity of *cultures* and *values*.[4]

Next, we test variation between nearest neighbors confirming Hellrich and Hahn's 2016a observation about the instability of nearest neighbors. Out of 25 nearest neighbors, only 2-5 are shared across all model types (an example is shown in Figure 3). However, these shared neighbors do confirm the initial observation about the changing meaning of *racial* and *cultural* (as presented in Tables 5 and 6).

In addition to differences between model algorithms, we also expect differences between SGNS models trained on the same corpus but with different initializations. We trained three SGNS models[5] for the COHA slices representative of the 1900s, 1950s and 1990s and compared the 25 nearest neighbors of *racial*. When considering the differences in the top 25 nearest neighbors of *racial* in the SGNS model trained on this comparatively small corpus, it can be seen that the number of shared neighbors between all three models ranges between 11 and 18 (Table 7). This means that as much as 14 out of 25 nearest neighbors vary depending on the three initializations, showing that drawing conclusions based on artefacts is indeed a risk. The number of shared neighbors increases with the size of the underlying subcorpus.

In order to gain deeper insight into the variation displayed by nearest neighbors, we examine the difference in rank of a specific word across various models. For instance, if *language* were ranked 5th closest in the model initialized with init1 and 15th closest in the model initialized with init2, the rank difference would be 10. Table 8 presents the average rank differences for the top 25 nearest neighbors of *racial* for each model pair. The average differences range from as high as 49 ranks difference in the smallest corpus to 6.24 in the largest corpus, again indicating higher stability with an increasing corpus size.

These results confirm Hellrich and Hahn's 2016a observation that even models trained on the same data created with the same method can lead to different conclusions depending on their initialization. As the initialization vectors are chosen randomly, there is a high risk of drawing conclusions due to artefacts rather than actual changes in the data when relying on a single, prediction-based model, in particular when trained on a small corpus. These risks can be reduced by creating multiple models and measuring the degree of difference between them. Based on the differences in rank of nearest neighbors, a larger environment can be studied to verify which changes are stable across models and larger than variations caused by artefacts.

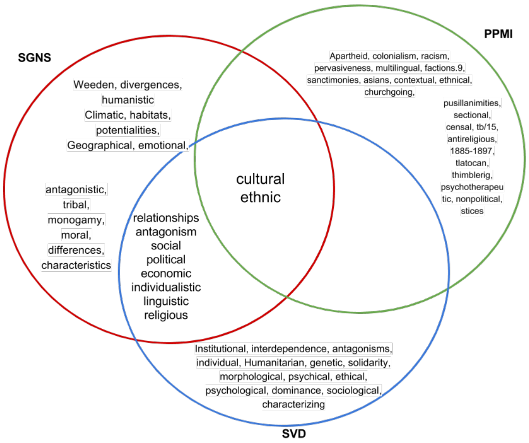


Figure 3: Nearest neighbors of *racial* in 1900 in different models created with the ngrams-corpus.

5.2 Control Words

Observations that hold across different models can still be a result of a bias or artefact in the

[4] In the experiments, equivalent part of speech (e.g. noun - noun) and number (e.g. plural - plural) have been chosen for investigating changes in word pairs.

[5] To train these models, we used a modified version of the code used by Levy et al. (2015) allowing us to fix the initialization vectors. We preprocessed the corpus with our own scripts, which may be slightly different from the preprocessing used by Hamilton et al. (2016).

1900	1950	1990
cultural ethnic	stereotypes ethnic backgrounds discrimination	discrimination segregation

Table 5: Nearest neighbors of *racial* shared across all three models in the n-gram corpus.

1900	1950	1990
racial morphological economic	socio racial social backgrounds ethnic	socio ethno

Table 6: Nearest neighbors of *cultural* shared across all three models in the n-gram corpus.

decades	**1900**	**1950**	**1990**
million tokens	25.7	29.0	33.2
init1-init2	15 (0.60)	15 (0.60)	20 (0.80)
init1-init3	16 (0.64)	18 (0.72)	20 (0.80)
init2-init3	16 (0.60)	16 (0.64)	19 (0.76)
init1-init2-init3	11 (0.44)	14 (0.56)	18 (0.72)

Table 7: Number of shared top 25 nearest neighbors of *racial* in the models created with three different initializations on the same decades of COHA.

decades	**1900**	**1950**	**1990**
million tokens	25.7	29.0	33.2
init1-init2	47.08	31.92	6.24
init1-init3	27.04	31.00	7.20
init2-init1	22.60	13.32	7.68
init2-init3	22.32	33.48	8.96
init3-init1	35.16	13.28	14.12
init3-init2	49.00	26.52	12.72

Table 8: Average differences in rank between the top 25 nearest neighbors of *racial* in the models created with three different initializations on the same decades of COHA.

'naive'	**data**	**models**	**control**
nn of *racial* indicate shift towards meta-discourse	yes	yes	n.a
cultures ⟷ *values*	yes	yes	yes
races ⟷ *immigrants*	no	partly	no
cultural ⟷ *superior*	no	yes (SVD), data sparsity (PPMI)	yes in n-grams, no in COHA
cultural ⟷ *inferior*	no	yes (SVD), no (PPMI)	yes

Table 9: Summary of results in line with the hypotheses in the 'naive' set-up.

data. Control words can potentially reveal such an underlying cause. If observations are indicative of changes in the relation between these specific words, control words should not reveal similar changes. To illustrate the insights that can result from such a test, we show the outcome of comparing *immigrants* and *races* in the COHA corpus in Figure 4. In this case, the control words may yield insights in addition to calling into question an apparent change in the usage of the word *immigrant*. It may have led to a new insight, namely that the actual change might lie in how the general concept of "People" relates to *races*, as the neutral control word *nurse* shows a highly similar pattern to the other social group labels. This outcome calls for further investigations to try and establish whether this is a pattern related to biological race, to *race* in the sense of speeding contest or to a difference in which one of these meanings occurs more frequently.

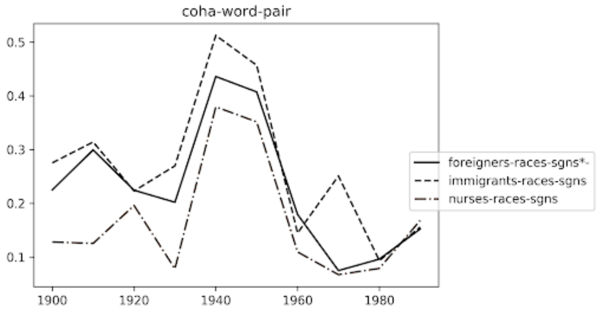


Figure 4: Changes in the cosine similarities between races and words representing social groups.

Overall, the results from the control experiments show that only a handful of the hypotheses were confirmed by all methods. Tables 9 and 10 provide an overview of the final outcome of our experiments in the different settings used to control for instability.

6 General Guidelines

Our experiments have shown that different models created from the same data do not always provide the same answers to our hypotheses. This outcome confirms the risk of naively applying distri-

‘naive’	**data**	**models**	**control**
nn of *racial - different*	no	no	no
racial - different ←→ values	yes	no	yes

Table 10: Summary of results contradicting the hypotheses in the ‘naive’ set-up.

butional semantic technologies to explore conceptual change. In particular when they seem plausible, there is a risk that results based on artefacts are presented as valid observations. Based on the outcome of this study, we propose the following guidelines for studying conceptual change using distributional semantics:

1. Define a wide range of verifiable hypotheses to study the overall question before diving into actual changes.
2. Compare the outcome of multiple models. Count-based models directly reflect the distribution in the data, but can be influenced by word frequency. When using predictive models, test variations with different initializations and different ordering of examples.
3. Adapt the range of nearest-neighbors based on the variation in rank across models to ensure that changes are indeed changes in distribution and not due to random artefacts of a predictive model.
4. Use control words that should not exhibit the same change to further verify your hypotheses. Ideal control words are close to those from the hypotheses, but lack the property that is supposed to have triggered the change (e.g. descriptions of racial groups vs. other descriptions of groups of people).

In addition, properties of the data (balance and size) should be taken into consideration. Control words can capture some of the problems that may be introduced by the data, but not all. Additional insights may be obtained by running verification experiments with shuffled and synchronic corpora as done in Dubossarsky et al. (2017).

7 Discussion and conclusion

Computational linguistics research has shown that distributional semantic models can be used to detect linguistic shifts (Hamilton et al., 2016), but has also shown that (a) not all observed changes are actual shifts (Hamilton et al., 2016; Dubossarsky et al., 2017) and that (b) predictive models can yield unstable results (Hellrich and Hahn, 2016a). We investigated the implications of this for a digital humanities use case: the concept “Racism”. Though the main insights from social science were confirmed by our study, most results turned out to be unstable.

A possible explanation is that our selection of words and relations is not representative of the actual conceptual system. As non-experts in the field of race studies, we selected the words and relations we investigated to the best of our abilities using existing literature. An interdisciplinary team might have proposed sounder hypotheses that would have been consistently confirmed. However, this does not undermine the, in our opinion, most important finding of this work. A standard, seemingly sound, experimental setup originally confirmed five hypotheses and showed clear patterns in nearest neighbors. Only two results could be reproduced by alternative methods and just 2-5 out of 25 nearest neighbors overlapped across all models. Furthermore, considerable variation was observed in the nearest neighbors of *racial* of models resulting from the same architecture and corpus with fixed different random initializations. Moreover, it should be noted that the impact of the order in which word-context pairs are considered by a prediction-based model has an impact on the results as well (Hellrich and Hahn, 2016b; Antoniak and Mimno, 2018). This variation has not been explored in this paper.

At this point, it is not possible to determine whether differences between models are due to random factors in prediction-based models, frequency effects in count-based models or a combination of both. However, the proposed checks and, in particular, investigations of the impact of random factors and patterns observed by control words provide a first step towards determining which results are artefacts and which are not. Identifying methods for answering this question is an important task for future work. We propose combining the guidelines resulting from this paper with the kind of experiments carried out by Dubossarsky et al. (2017) as a first next step.

This main contribution of this study is that it shows the risks of applying methods that work for specific examples and data to new use cases. It is tempting to assume that the method works when it provides an expected outcome or, otherwise, an outcome that can easily be explained. At this point, the relation between linguistic data and resulting semantic models is not understood well enough to draw conclusions from diachronic comparisons. Until we have more profound knowledge about the interpretation of shifts, conclusions about conceptual change should be drawn with care and verified through multiple means.

Acknowledgments

This research is funded by the PhD in the Humanities Grant provided by the Netherlands Organization of Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO) PGW.17.041 awarded to Pia Sommerauer and NWO VENI grant 275-89-029 awarded to Antske Fokkens. We would like to Minh Le for providing the adapted version of word2vec that allows us to select a specific initialization model and we thank anonymous reviewers for their comments. All remaining errors are our own.

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association of Computational Linguistics*, 6:107–119.

Martha Augoustinos and Danielle Every. 2007. The language of race and prejudice a discourse of denial, reason, and liberal-practical politics. *Journal of Language and Social Psychology*, 26(2):123–141.

Martin Barker. 1981. *The new racism: conservatives and the ideology of the tribe*. Junction Books.

Arianna Betti and Hein van den Berg. 2014. Modelling the history of ideas. *British Journal for the History of Philosophy*, 22(4):812–835.

Johannes Bjerva and Raf Praet. 2015. Word embeddings pointing the way for late antiquity. In *LaTeCH@ ACL*, pages 53–57.

Mark Davies. 2002. *The Corpus of Historical American English (COHA): 400 million words, 1810-2009*.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *NetWordS*, pages 66–70.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1136–1145.

Steffen Eger and Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 52–58.

John Rupert Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1489–1501.

Johannes Hellrich and Udo Hahn. 2016a. Bad companyneighborhoods in neural embedding spaces considered harmful. In *COLING (16)*, page 27852796.

Johannes Hellrich and Udo Hahn. 2016b. An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. In *LaTeCH 2016Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities@ ACL*, pages 111–117.

Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1191–1200. ACM.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *ACL 2014*, page 61.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. ACM.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

Andrey Borisovich Kutuzov, Elizaveta Kuzmenko, and Anna Marakasova. 2016. Exploration of register-dependent lexical semantics using word embeddings. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 26–34.

Alana Lentin. 2005. Replacing race, historicizing culturein multiculturalism. *Patterns of prejudice*, 39(4):379–396.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Carlos Martinez-Ortiz, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul, and Joris Van Eijnatten. 2016. Design and implementation of shico: Visualising shifting concepts over time. In *HistoInformatics 2016*, volume 1632, pages 11–19.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, volume 13, pages 746–751.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. *arXiv preprint arXiv:1405.4392*.

Ann Morning. 2009. Toward a sociology of racial conceptualization for the 21 st century. *Social Forces*, 87(3):1167–1192.

Michael A Omi. 2001. The changing meaning of race. *America becoming: Racial trends and their consequences*, 1:243–263.

Matthias Orlikowski, Matthias Hartung, and Philipp Cimiano. 2018. Learning diachronic analogies to analyze concept change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11.

Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Unni Wikan. 1999. Culture: A new concept of race. *Social Anthropology*, 7(01):57–64.

Howard Winant. 1998. Racism today: Continuity and change in the post-civil rights era. *Ethnic and Racial Studies*, 21(4):755–766.

Gerhard Wohlgenannt, Ariadna Barinova, Dmitry Ilvovsky, and Ekaterina Chernyak. 2019. Creation and evaluation of datasets for distributional semantics tasks in the digital humanities domain. *arXiv preprint arXiv:1903.02671*.

Erjia Yan and Yongjun Zhu. 2018. Tracking word semantic change in biomedical literature. *International journal of medical informatics*, 109:76–86.

A Detailed overview of hypotheses and outcomes

word1	word2	Hypothesis	Coha-sgns	Ngrams-sgns
racial	cultural	closer	-	-
racial	superior	apart	-	-
racial	inferior	apart	apart	-
racial	blacks	apart	-	-
racial	whites	apart	apart	closer
racial	marriage	apart	-	closer
racial	relationships	apart	-	-
racial	genetics	apart	OOV	OOV
racial	nigger	apart	closer	closer
racial	yankee	apart	-	-
racial	gypsy	apart	-	-
cultural	superior	apart	closer	apart
cultural	inferior	apart	-	apart
cultural	blacks	apart	-	closer
cultural	whites	apart	-	closer
cultural	marriage	apart	-	apart
cultural	relationships	apart	-	-
cultural	genetics	apart	OOV	OOV
cultural	nigger	apart	closer	-
cultural	yankee	apart	-	-
cultural	gypsy	apart	-	-
racial	immigrant	closer	-	apart
racial	foreigner	closer	apart	-
racial	national	closer	-	apart
racial	Turks	closer	OOV	OOV
racial	Arabs	closer	-	-
racial	Jews	closer	apart	-
racial	religious	closer	closer	-
racial	linguistic	closer	-	-
racial	values	closer	apart	closer
racial	attitudes	closer	-	apart
racial	beliefs	closer	-	apart
racial	historic	closer	apart	-
racial	different	closer	-	-
cultural	immigrant	closer	-	-
cultural	foreigner	closer	-	-
cultural	national	closer	closer	-
cultural	Turks	closer	-	-
cultural	Arabs	closer	-	-
cultural	Jews	closer	-	-
cultural	religious	closer	closer	-
cultural	linguistic	closer	-	closer
cultural	values	closer	closer	closer
cultural	attitudes	closer	-	-
cultural	beliefs	closer	-	-
cultural	historic	closer	-	-
cultural	different	closer	closer	-

Table 11: Overview of hypothesized changes and results in of the SGNS model in COHA and the google n-grams. The forms of *racial* and *cultural* have been adapted to match word2 in part of speech and number. *closer* indicates a significant change towards each other and *apart* a significant increase in distance, - means no significant change