Measuring diachronic evolution of evaluative adjectives with word embeddings: the case for English, Norwegian, and Russian

Julia Rodina[1], **Daria Bakshandaeva**[1], **Vadim Fomin**[1],
Andrey Kutuzov[2], **Samia Touileb**[2], and **Erik Velldal**[2]

[1]National Research University Higher School of Economics, Moscow, Russia
[2]Language Technology Group, University of Oslo, Oslo, Norway
{julia.rodina97,dbakshandaeva,wadimiusz}@gmail.com
{andreku,samiat,erikve}@ifi.uio.no

Abstract

We measure the intensity of diachronic semantic shifts in adjectives in English, Norwegian and Russian across 5 decades. This is done in order to test the hypothesis that evaluative adjectives are more prone to temporal semantic change. To this end, 6 different methods of quantifying semantic change are used.

Frequency-controlled experimental results show that, depending on the particular method, evaluative adjectives either do not differ from other types of adjectives in terms of semantic change or appear to actually be less prone to shifting (particularly, to 'jitter'-type shifting). Thus, in spite of many well-known examples of semantically changing evaluative adjectives (like 'terrific' or 'incredible'), it seems that such cases are not specific to this particular type of words.

1 Introduction

Words change their meaning over time. It has become widespread recently to trace such shifts using word embedding models (that is, using contextual cues from raw corpora). However, most of this research is centred on the English language, and focuses on *nouns* specifically. In this paper, we work with 3 different languages (English, Norwegian and Russian), and focus our attention on *adjectives*.

Particularly, we aim to test empirically whether *evaluative adjectives* are more susceptible to diachronic semantic shifts than other types of adjectives. Evaluative adjectives are defined as those which describe object qualities from the subjective point of view of the speakers, expressing their opinions about the object being described. Typical English examples are '*good*', '*bad*' or '*brilliant*'.

Sometimes, adjectives can become evaluative in the course of semantic shifts happening across time: consider the history of the English word *monumental* from the 60s to the 2000s (Figure 1)[1] or how the word *sick* slowly acquires a (colloquial) evaluative sense ('*That's sick, dude!*') as described in Mitra et al. (2014). On the other hand, intuitively, evaluative adjectives are naturally prone to amelioration and pejoration as major types of diachronic semantic shifts. One can immediately recall, for example, the English words *incredible* and *terrific* which underwent amelioration and started to denote positive instead of negative qualities.

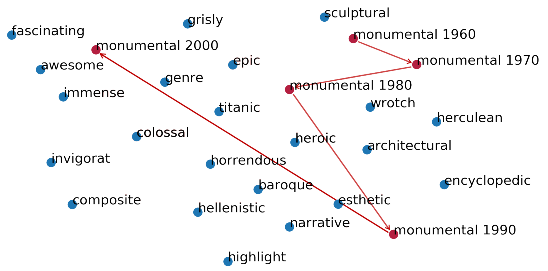

Figure 1: Alterations in meaning of the English adjective '*monumental*': from *sculptures* in the sixties to *awesome* in the 2000s

But are these words only isolated hand-picked examples, or is there a general trend in human languages which makes evaluative adjectives change more intensely over time? In this paper, we try to answer this question. Section 2 puts this work in the context of previous research. In section 3, we describe the corpora and word lists we relied upon. Our experiments are described in 4. In sections 5 and 6 we outline the limitations of the presented research, our plans for the future, and conclude.

[1] See Appendix A for details on visualisation

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 202–209
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

2 Related work

The nature of semantic change processes has always been of special interest to linguistics. This interest started at least as early as in Bréal (1883) who asserted the intellectual (cognitive) laws of semantic change as opposed to 'natural' ones. Later, Bloomfield (1933) proposed a popular categorisation of semantic shifts into classes. Further on, the academic community tried to develop a theoretical understanding of reasons behind semantic shifts, and to refine their classification (Meillet, 1974).

Moving on to specific types of semantic shifts, amelioration (acquiring more positive sentiment) and pejoration (acquiring more negative sentiment) were studied in Borkowska and Kleparski (2007), who mentioned these types to be one of the strongest and most wide-spread.

As the amount of language data available to computational linguistics increased,[2] the focus of research interest moved from theoretical reasoning about the nature of semantic shifts to more empirical approaches, mainly based on corpus-based analysis (see Michel et al. (2011) and Jatowt and Duh (2014), among many others).

Recently, the usage of pre-trained word embeddings (Bengio et al., 2003; Mikolov et al., 2013a) has become widespread in the publications related to diachronic semantic shifts (Kim et al., 2014; Hamilton et al., 2016c; Liao and Cheng, 2016; Kutuzov et al., 2017b,a; Rosenfeld and Erk, 2018). The main reason for this is the powerful abilities of such approaches to model word meaning based solely on non-annotated corpora. Additionally, vector representations of words allow for easy calculation of their similarities and changes. The baseline method here consists of simply training embedding models on the texts created in different time periods, and then comparing the vector representations for the same words. For further information on the current state of the field, see Kutuzov et al. (2018) and Tang (2018).

One of the difficulties brought by these approaches is the necessity to somehow 'align' the vector spaces trained on different time bins (time periods). A variety of methods have been proposed to overcome this. They include initialising the models for each time bin with the weights from models trained on the previous time bin ('incremental training') (Kim et al., 2014); Procrustes alignment of independent embedding models (Hamilton et al., 2016c); dynamic models trained across all time bins at once (Bamler and Mandt, 2017; Yao et al., 2018; Rosenfeld and Erk, 2018); Global Anchors (measuring the vectors of words' similarities to other words) (Yin et al., 2018), etc. In this paper, we employ Procrustes alignment and the Global Anchors methods, applying them to the task of measuring the speed of semantic shifts of evaluative adjectives across time.

An important publication related to our work is Hamilton et al. (2016a). In it, the authors induce historical sentiment lexicons from English corpora (using word embeddings, among other methods). They further show that amelioration and pejoration do occur on a massive scale: many evaluative adjectives in English have completely switched their sentiment during the last 150 years. We extend this work by studying not only sentiment changes, but semantic shifts in evaluative adjectives in general. Additionally, we analyse data from 3 languages (English, Norwegian and Russian). However, we focus on a more narrow time span: only the decades from 1960s to 2000s.

3 Data

In this section, we describe our data: the corpora employed to train word embedding models, and the sentiment lexicons serving as the source of evaluative adjectives.

3.1 Corpora

For the purposes of our research, we employed corpora in three languages, selecting texts which were created during the five decades from 1960s to 2000s. We lemmatized (it was especially important for Russian with its rich morphology) and POS-tagged all the corpora ourselves, using the corresponding UDPipe models (Straka and Straková, 2017).

For *English* data, we used The Corpus of Historical American English (COHA).[3] This is a corpus of English texts annotated with creation dates and balanced by genres. It is composed of fiction, magazine and newspaper articles, as well as non-fiction texts.

[2] For example, the Google Ngrams (`https://books.google.com/ngrams`) service stimulated diachronic research of texts and language greatly.

[3] `https://www.english-corpora.org/coha/`

Decade	**English**	**Norwegian**	**Russian**
1960s	12	6	10
1970s	12	21	10
1980s	13	25.5	9
1990s	14.5	40.5	20
2000s	15	21	39.5

Table 1: Corpora sizes (in millions of words)

For *Norwegian* data, we used the NBdigital corpus.[4] It contains texts in Norwegian Bokmål from the National Library of Norway's collection of free texts, obtained by OCR processing (only texts with the OCR confidence higher than 0.9). These texts were mainly produced by various public institutions.

For *Russian* data, we used the Russian National Corpus (RNC).[5] It includes a wide variety of genres of written and spoken language, such as non-translated works of fiction, memoirs, essays, journalistic works, scientific and popular scientific literature, public speeches, letters, diaries, documents, etc. It is important that the RNC is also rigorously balanced across genres and types of texts.

Table 1 lists the corpora sizes for each decade under consideration.

3.2 Word embeddings

Continuous bag-of-words (CBOW) embedding models (Mikolov et al., 2013b) were trained on each decade for each of the three languages. All the models share the same set of hyperparameters: vector size 300, symmetric context window size 3, and 10 iterations over the corpus. We discarded all the words which occurred less than 5 times in the training corpus, and additionally limited the maximum vocabulary size to be 100 000, more or less following the hyperparameters from Kutuzov et al. (2017a). The models are made available via the NLPL word vector repository[6] (Fares et al., 2017).

3.3 Evaluative adjectives lists

In order to find out whether evaluative adjective are more prone to diachronic semantic shifts, we need an authoritative source providing us with a list of such adjectives, more than only several words in size. Unfortunately, even for English such a list is hard to find in the published works, and the same is true for Norwegian and Russian. For this reason, we turned to sentiment lexicons: lists of positive and negative words widely used for the purposes of automatic sentiment analysis. The ratio is that such words are almost always evaluative by definition. Below we describe these lexicons for each of the three languages under analysis.

The lists for *English* and *Norwegian* come from the same source. The English lexicon is a general sentiment lexicon composed of a positive and a negative lexicon. These were created by assigning the positive and negative labels using a WordNet-based bootstrapping approach (Hu and Liu, 2004)[7]. We thereafter automatically translated (from English to Norwegian) these positive and negative sentiment lexicons. The translations were manually checked, and corrected when necessary. Furthermore, if an English word had several senses that could be translated into different Norwegian words, these were added to the translations. We have omitted all multi-word expressions, and only kept single word translations. This resulted in a collection of 3961 negative and 1646 positive Norwegian words. The original English lexicons contained 4783 negative and 2006 positive words. We did not investigate rigorously to what extent the translated lexicon is representative of the Norwegian language, but we believe that it is representative enough, since it is a general lexicon equivalent to its original English counterpart, and because the Norwegian list was checked manually to filter out non-evaluative adjectives.

The Norwegian lexical resource SCARRIE[8], a full-form lexicon, was used to identify which of the Norwegian translations were adjectives. Once these Norwegian adjectives were identified, we selected only the English words that had a Norwegian adjective as translation. Subsequently, we used the WordNet (Miller, 1995) to identify which of the selected English words were actually adjectives. If an English word was not identified as an adjective, we used WordNet to find its adjective form by analysing the derivationally related forms of its lemma. If no such form could be found, then the English word was removed from our list. Both lists were thereafter lemmatized and manu-

[4] https://www.nb.no/sprakbanken/show?serial=oai:nb.no:sbr-43&lang=en

[5] http://ruscorpora.ru/en/

[6] http://vectors.nlpl.eu/repository/

[7] Available at https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[8] https://www.nb.no/sprakbanken/show?serial=sbr-9&lang=nb

ally filtered to remove non-evaluative adjectives. This resulted in 2250 English adjectives and 1939 Norwegian adjectives.

We borrowed *Russian* evaluative adjectives from *RuSentiLex* (Loukachevitch and Levchik, 2016): a list of sentiment-related words and expressions. There are three types of entries in *RuSentiLex*, depending on their source: 'opinion', 'feeling' and 'fact' (words or expressions that do not express an opinion of the author, but have a positive or negative connotation). Also, each entry is labelled with its part of speech, lemmatized form and polarity, which can be positive, negative, neutral or positive/negative for strong context-dependent semantic orientation. Polysemous words have separate entries for different senses. The current version of the lexicon contains more than 12 thousand words and expressions, which were semi-automatically obtained from existing domain-oriented sentiment vocabularies (initial list), news articles (words with connotations) and Twitter (slang and curse words). For this research we used only one-word adjectives labelled with the 'opinion' source. Since differences in the sentiment and polarity of polysemous words are not taken into account in this paper, repeated entries have been removed. In total, there are 2435 Russian evaluative adjectives.

After acquiring the lists of evaluative adjectives and training word embedding models on the texts created for each decade under analysis, we were able to move on to the experiments.

4 Experiments

Our general aim is to measure the speed of temporal semantic shifts in evaluative adjectives compared to all other adjective types. This is necessary to confirm or reject the hypothesis that evaluative adjectives are less stable than other words of the same part of speech. We want to find evidence across all three languages under analysis. We also would like to control for frequency and to exclude its influence on the results, since it is known that word frequency often correlates with the speed of semantic change (Hamilton et al., 2016c) [9].

We measure the speed of semantic changes using a variety of methods:

1. *Jaccard distance* (Jaccard, 1901) between sets of 10 nearest neighbours of one word (by cosine distance) in two embedding models;
2. *Procrustes alignment* (Hamilton et al., 2016c): the models' vector spaces are first aligned using an SVD-based orthogonal transformation, and then cosine distance is calculated between one word's vectors in two transformed models;
3. *Global Anchors* (Yin et al., 2018): here, the degree of semantic change is defined as the cosine distance between the vectors of a word's cosine similarities to all other words in the intersection of two models' vocabularies ('anchors').

The aforementioned methods measure the distance between the meanings of one word in two different embedding models. However, our data includes five models (trained on five consequent decades from 1960s to 2000s). In order to quantify the speed of semantic change across the whole time span, two techniques were used:

1. *Mean distances*: simple mean between the 4 pairwise distances ('60s to 70s', '70s to 80s', '80s to 90s', and '90s to 2000s'). It measures the degree of 'semantic jitter' that the word undergoes: it is not necessarily a steady movement into one direction, but can instead be fluctuations around one centre point (points).
2. *Mean deltas from the 60s*: here, at each decade we calculate the distance of the current word representation to its representation in the 60s (the initial point of our time span). If the distance increased, one point is added to the word's score, if the distance decreased, one point is subtracted. Then, the average score is calculated for each word. The rationale behind this is to measure how steady the shift in meaning is from the initial point for a given word. The score here will be low for the words which fluctuate but do not really substantially change their semantics. At the same time, it will be high for consistent cases (like, for example, the English adjective '*solid*' steadily moving toward denoting not only qualities of materials, but also generally being of good quality). See, e.g., Figure 2 for an example of how a word can first move away from the original meaning, but then start to slowly return back.

[9] Note, however, that this was disputed in Dubossarsky et al. (2017).

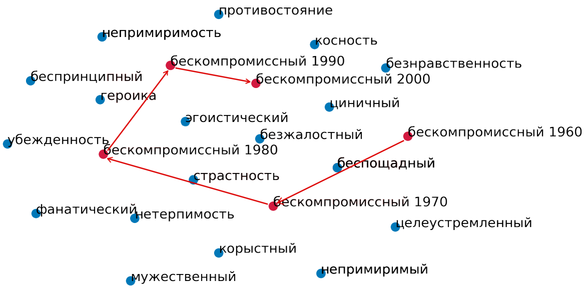

Figure 2: Alterations in meaning of the Russian adjective 'бескомпромиссный' (*uncompromising*): from *ruthless* over *fanatical*, *passion*, later *conviction*, *heroic* to *intransigence*, *confrontation*

Both *mean distances* and *mean deltas from the 60s* can be used with any method of measuring semantic change of the 3 described above, thus overall we have 6 scores to assign to each word in our word lists.

Note that we have *two* word lists for each language: the one with *evaluative* adjectives (extracted from sentiment lexicons) and another with what we will refer to as *fillers*: that is, simply all other adjectives present in the vocabularies of all five models for the current language. We compare the semantic change speed scores of the first list to those in the second one. If the average values differ with the Welch's T-test p-value not exceeding 0.1[10], we conclude that one type of adjectives is more subject to diachronic semantic change than the other, and report the t-statistics of the difference between the averages. If the p-value exceeds the 0.1 threshold, we conclude there is no difference between two lists, and report it as 0 (full anabridged tables available at https://github.com/ltgoslo/diachronic_multiling_adjectives/tree/master/full_tables).

Table 2 presents the results calculated this way. Positive t-statistic values mean that evaluative adjectives change faster than other types of adjectives, according to particular metrics; negative values mean they change slower. We also report the number of filler adjectives ('# fillers') for each language.

Method	**English**	**Norwegian**	**Russian**
# fillers	8994	3989	7535
Freq diff	0.00001	0.00003	0.00001
	Mean pairwise distances		
Jaccard	-11.08	-4	-15.05
Procrustes	-15.52	-5.04	-12.01
GlobAnchors	11.91	-4.40	12.62
	Mean deltas from 1960s		
Jaccard	3.28	0	0
Procrustes	2.98	0	3.92
GlobAnchors	3.57	3.24	3.11

Table 2: Difference in the intensity of semantic shifts between evaluative adjectives and fillers. Positive values correspond to evaluatives changing significantly faster, and vice versa.

As can be seen, across all languages, evaluative adjectives seem to fluctuate less (*mean pairwise distances*), as measured by all methods, except for Global Anchors applied to English and Russian. At the same time, the majority of methods agree that evaluative adjectives are more likely to steady shift in one direction, farther and farther away from the original meaning (as measured by *mean deltas from the 60s*). This is less expressed for Norwegian (with Jaccard and Global Anchors methods, the difference between two types of adjectives was not significant).

However, these values are potentially problematic. As already mentioned, the speed of semantic change can correlate with word frequencies. The 'Freq diff' line in the table 2 shows the difference between average word frequencies in both word lists (expressed as word probabilities relative to corpora sizes). All these values are statistically significant and positive: this means that evaluative adjectives are on average more frequent than other adjectives.

Table 3 proves that there are indeed statistically significant correlations between word frequencies and all our methods for measuring the intensity of temporal semantic shifts, across all languages. More frequent words consistently get *lower* scores from *mean distances*.[11] Vice versa, they get *higher* scores from the *mean deltas* technique, suggesting that frequent words are more prone to steady semantic shifting.

[10] The p-value threshold of 0.1 was used intentionally, instead of the more standard 0.05. We could as well use 0.05, and it wouldn't change the final results of the research (the original hypothesis would still be rejected). The reason behind choosing 0.1 was to be able to show that some differences in the speed of semantic change between evaluative adjectives and fillers can be found, but they are rare and fragile even with a very permissive p-value threshold.

[11] It seems to support the law of conformity from Hamilton et al. (2016c)

Method	English	Norwegian	Russian
	Mean distances		
Jaccard	-0.37	-0.33	-0.32
Procrustes	-0.19	-0.21	-0.17
GlobAnchors	0.29	-0.08	0.11
	Mean deltas from 1960s		
Jaccard	0.05	0.10	0.08
Procrustes	0.07	0.12	0.08
GlobAnchors	0.07	0.12	0.05

Table 3: Correlation of semantic change speed and normalised word frequency across all adjectives (evaluative and fillers). Positive values correspond to frequent words changing significantly faster, and vice versa.

Method	English	Norwegian	Russian
# fillers	1133	571	929
Freq diff	0	0	-0.00002
	Mean distances		
Jaccard	0	-1.68	-2.54
Procrustes	-4.77	-3.24	-5.03
GlobAnchors	-3.70	-4.07	0
	Mean deltas from the 1960s		
Jaccard	0	0	-2.44
Procrustes	0	2.94	0
GlobAnchors	0	0	-1.79

Table 4: Difference in the intensity of semantic shifts between evaluative adjectives and fillers (frequency > 100). Positive values correspond to evaluatives changing significantly faster, and vice versa.

To get rid of the influence of the frequency factor in comparing evaluative and non-evaluative adjectives, we have to make the average frequencies of both lists more like each other. Since we observed that evaluative adjectives are more frequent, we decided to use the *frequency threshold*. All adjectives with corpus frequency in at least one decade lesser than the threshold (which is a hyperparameter) were removed from the word lists (both evaluative adjectives and fillers) [12]. This allowed us to get rid of low-frequency long-tail and make both lists to better fit each other in terms of frequency. In the table 4, we report the results using the threshold of 100.

The number of fillers has naturally declined. Also, the 'Freq diff' line shows that this way we managed to eliminate any statistically significant difference between evaluative and non-evaluative word lists for English and Norwegian. For Russian data, the situation has reversed: now evaluative adjectives are on average *less* frequent. Interestingly, the overall results for the 'mean distances' methods did not change or even became more expressed. Even when controlled for frequency, evaluative adjectives seem to be less prone to 'fluctuating' semantic shifts. Thus, to some extent they are more semantically stable than other adjectives. This makes us reject the initial hypothesis about them being less stable.

Note that for the *mean deltas* technique, filtering out the low-frequency words led to the differences between evaluative and non-evaluative adjectives losing their statistical significance in almost all combinations of languages and methods. Thus, we cannot prove any specificity of evaluative adjectives with respect to the 'steadiness' of diachronic semantic changes.

5 Limitations and future work

First of all, sentiment lexicons as sources of 'evaluative adjectives' are by all means only proxies. It is quite probable that there are evaluative adjectives beyond sentiment lexicons, and vice versa. In the future, we plan to refine our datasets and probably come up with more linguistically justified word lists.

Although we used the well-known methods of measuring semantic shifts across word embedding models, there is still a need to evaluate the methods themselves. One option here it to use the *SentProp* historical sentiment datasets from Hamilton et al. (2016b). These datasets are created automatically, but still this sanity check could allow us to find out which of the algorithms produces results better correlated with the output of other systems. At the same time, it is known that distributional models can have a hard time handling the differences between antonyms, and those constitute a significant part of diachronic changes in *SentProp* (cf. '*incredible*' changing it sentiment from negative to positive in the last 40 years). There is an ample room for further research here.

Note also that the interplay between semantic shift detection methods and word frequencies is quite complex, and there is still a room to investigation. We didn't analyse it deeply, so we can-

[12] We did not down-sample the evaluative adjectives instead, since they are the main focus of our research, and we did not want to reduce their number (not huge to begin with).

not exclude the possibility that the results could change if controlling for other related factors.

6 Conclusion

We measured the intensity of diachronic semantic shifts in adjectives across 3 languages (English, Norwegian and Russian) and 5 decades (60s, 70s, 80s, 90s, 2000s), to test whether evaluative adjectives change faster (or more intensely) than other adjectives.

Our results show that, contradictory to the initial hypothesis, evaluative adjectives change over time *less* intensely (statistically significant at $p < 0.1$), if we measure change as the mean of pairwise differences between successive decades, and not as a steady 'movement' in one particular direction. This is not an artefact of frequency, since we observe the same behaviour when controlling for word frequencies.

At the same time, when measuring the probability of steady 'moving away' from an original meaning across time, evaluative adjectives *do not differ from other adjectives at all* (at least on any statistically significant level).

To sum up, it seems that evaluative words (in our case, adjectives) are not more prone to semantic shifts than other word types. Vice versa, under some circumstances, they can be even more stable than their counterparts, with this observation holding across languages and methods of semantic shifts tracing.

Our diachronic embedding models, word lists and code can be found at https://github.com/ltgoslo/diachronic_multiling_adjectives.

Acknowledgements

Parts of this work has been carried out in context of the SANT project, as funded by the Research Council of Norway (project number 270908).

References

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the International Conference on Machine Learning*, pages 380–389, Sydney, Australia.

Yoshua Bengio, Rejean Ducharme, and Pascal Vincent. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Leonard Bloomfield. 1933. *Language*. Allen & Unwin.

Paulina Borkowska and Grzegorz Kleparski. 2007. It befalls words to fall down: pejoration as a type of semantic change. In *Studia Anglica Resoviensia*, volume 47(4), pages 33–50.

Michel Bréal. 1883. Les lois intellectuelles du langage: fragment de sémantique. *Annuaire de l'Assocaition pour l'encouragement des études grecques en France*, 17:132–142.

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 1147–1156, Copenhagen, Denmark.

Murhaf Fares, Andrey Kutuzov, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276. Association for Computational Linguistics.

William Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016a. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas.

William Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016b. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016c. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, Berlin, Germany.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Paul Jaccard. 1901. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 229–238, Piscataway, NJ, USA. IEEE Press.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the 52nd Annual Meeting of the Association* *for Computational Linguistics*, pages 61–65, Baltimore, USA.

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. Association for Computational Linguistics.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017a. Temporal dynamics of semantic relations in word embeddings: an application to predicting armed conflict participants. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1824–1829, Copenhagen, Denmark.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017b. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop at ACL 2017*, pages 31–36, Vancouver, Canada.

Xuanyi Liao and Guang Cheng. 2016. Analysing the semantic change based on word embedding. In *Natural Language Understanding and Intelligent Applications*, pages 213–223. Springer International Publishing.

Natalia Loukachevitch and Anatolii Levchik. 2016. Creating a general Russian sentiment lexicon. In *Proceedings of Language Resources and Evaluation Conference (LREC-2016)*, pages 1171–1176.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85.

Antoine Meillet. 1974. Wie die wörter ihre bedeutung ändern. *G. Disner (ed.) Zur Theorie der Sprachveränderung*, pages 19–67.

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, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119.

George Miller. 1995. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

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. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1020–1029, Baltimore, Maryland.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 474–484, New Orleans, Louisiana, USA.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99. Association for Computational Linguistics.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 673–681, Marina Del Rey, CA, USA.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The global anchor method for quantifying linguistic shifts and domain adaptation. In *Advances in Neural Information Processing Systems*, pages 9433–9444.

A Visualization

Visualisation algorithm is based on the method described in (Hamilton et al., 2016c). To trace visually the movement of a given word in the semantic space we take a union of m most similar words for all periods that are of interest for us. Then t-SNE (Van der Maaten and Hinton, 2008), a technique for dimensionality reduction, is used to fit embeddings into two dimensional space: for m nearest neighbours, t-SNE embedding is found only on the most recent period (which represents the most recent meanings of these words), whereas for the word under consideration, embeddings from all time periods are taken into account. Procrustes alignment is preliminarily applied so that embeddings of the target word from all time bins are placed in common embedding space.