A Study on Word2Vec on a Historical Swedish Newspaper Corpus

Nina Tahmasebi

Språkbanken & Center for Digital Humanities,
University of Gothenburg, Sweden
nina.tahmasebi@gu.se

Abstract. Detecting word sense changes can be of great interest in the field of digital humanities. Thus far, most investigations and automatic methods have been developed and carried out on English text and most recent methods make use of word embeddings. This paper presents a study on using Word2Vec, a neural word embedding method, on a Swedish historical newspaper collection. Our study includes a set of 11 words and our focus is the quality and stability of the word vectors over time. We investigate whether a word embedding method like Word2Vec can be effectively used on texts where the volume and quality is limited.

1 Introduction

Automatic detection of word sense change has been investigated for the past decade or so, but has received increasing attention in recent years with (neural) word embeddings as a new way forward. There are many reasons why detecting word sense change is necessary; in addition to being interesting on its own (when and how a word changes its meaning(s)), it is also needed for understanding documents retrieved from historical corpora and for computationally detecting, for example, sentiments over time.

Previous methods for automatic detection of word sense change have included the comparison of context vectors, topic models and graph-based models as well as word embeddings. The topic modeling and graph-based methods aim to separate a word into its different senses and make predictions for a word based on its individual senses. The context based methods and lately the word embedding methods have made use of representations of the whole word rather than its senses. These methods typically detect changes in the main (dominant) sense of a word and cannot distinguish between stable senses and changing ones. The deficiency of word embedding models can be overcome by using one embedding per sense and then tracking these embeddings over time, allowing sense differentiations like one or more stable senses and one or more changing ones to capture the full picture. (More on related work in Section 4.)

Thus far, most, if not all, investigations into automatic detection of word sense change have focused on English texts for several reasons, the availability of large diachronic corpora being the most important one. Many (neural) embedding methods require large amounts of data and, therefore, the applicability

of these methods are limited for languages and time spans that do not have the required volume of digital data. This problem becomes even more acute if we wish to make use of sense-differentiated embeddings where there needs to be enough data for each sense of a word, thus increasing the data requirements.

In this paper, we will investigate the Word2Vec model [1] using the Swedish historical newspaper archive Kubhist [2]. We consider this a feasibility study on neural embeddings for the Kubhist material and, assuming the results show reasonable quality, a starting point for automatic word sense change detection on the basis of sense-differentiated word embeddings.

We make use of 11 words, nyhet 'news', tidning 'newspaper', politik 'politics', telefon 'telephone', telegraf 'telegraph', kvinna 'woman', man 'man', glad 'happy', retorik 'rethoric', resa 'travel' and musik 'music'. While some of these words represent rather stable concepts (*e.g.* news, happy) others represent new concepts (*e.g.* telegraph, telephone) and some have the potential to reveal interesting cultural changes (*e.g.* woman, rhetoric, travel). We begin by explaining our method and then analyze the results for some words. Tables of top k words not discussed in the paper can be found in the appendix. The full set of results (all years, all top 10 words) can be found in [3].

2 Method

We begin with the Kubhist data making use of years 1749-1925, (excluding Aftonbladet which was added later)[1]. The data can be found and investigated in Språkbanken's research tool Korp [4]. The 78 papers included in the corpus consist of 876 million tokens and close to 69 million sentences. Starting in 1845, there are over 5 million tokens per year and over 14 million tokens at most in year 1879. We lemmatize the data using the Korp infrastructure and replace each word with its lemma. We apply Word2Vec (W2V), which is a two-layer neural net out of the box using the Deeplearning4j (DL4J) package for Java [5].

We run the W2V models for each year of the dataset separately. Because vectors cannot be compared directly when trained on different corpora (they need to be projected onto the same space first) we make use of the words that are closest to a vector. That means, for each word w that we investigate, we print out the 10 words corresponding to the 10 closest vectors to the vector of w for a given year, i.e., the 10 most similar words to w. When all years are processed, we have a table for each word w, where each line corresponds to a year and contains the 10 closest words. Certain years will have no words because a vector could not be found corresponding to w, i.e., there was too little evidence for w in the corpus during that year.

To investigate these tables, we study their content closely, but we also make use of some statistics. We are mainly interested in how stable the vector spaces are. If there is word sense change, the vectors should be changing. However, far

[1] There is not sufficient data in all years for producing vectors. In addition, year 1758 is included for some words and not others and therefore we have chosen to exclude the year for all words.

word	avg. Jaccard (A)	avg. freq. (B)	corr(A,B)
telegraph	0.029	16.592	0.466
politics	0.042	31.816	0.472
news	0.048	33.018	0.475
woman	0.019	39.984	0.665
happy	0.019	56.242	0.182
telephone	0.032	62.127	0.461
music	0.179	68.681	0.607
travel	0.134	311.999	0.129
newspaper	0.191	381.541	0.327
man	0.105	2084.984	0.067

Table 1. Values for average Jaccard similarity, average normalized frequency and the Spearman correlation between the two, sorted on decreasing frequency.

from every change in the vector space corresponds to word sense change. Since radical sense change is relatively rare, we use the stability of the vector space as a quality measure of the vectors.

To measure the stability, we ask how many of the top 10 words that appear year after year and find this by calculating the Jaccard similarity between each pair of adjacent years. The Jaccard similarity measures, given two lists of words A and B, the overlap between A and B, divided by the number of unique items in both A and B. For example, A = {happy, smiling, glad} and B = {happy, joyful, cheerful, excited}, then the overlap of A and B is 1 (since they share the word *happy*) and there are 3 + 4 - 1 = 6 unique words. The Jaccard similarity is then 1/6 = 0.167. To be able to investigate how the Jaccard similarity changes over time, we plot the smoothened Jaccard similarity over time. The smoothing aims to make the graph simpler to investigate and is the average value of three years – year i, the one that is plotted, the preceding year i-1 and following year i+1. The exceptions are the first and the last years (1749 and 1925) where only two years are taken into account.

To put the Jaccard similarities into context, we also plot the normalized frequency of the word w from the corpus. The normalized frequencies are computed by Korp and are not smoothened. The correlation values between the Jaccard similarities and the normalized frequencies are calculated on the non-smoothened Jaccard similarities (while the smoothened ones are in the plots for visual reasons). Finally, we will provide tables for each word where the top 10 words can be viewed for certain years.

3 Results

We begin by noting that out of the 11 investigated words, one did not have any vector representations at all due to its low frequency in the corpus. The words retorik 'rhetoric' appears 78 times during the entire time span of Kubhist, which amounts to at most three occurrences for one year. That means, we have in total 10 words left for our investigation, however, in the case of 'woman', we make use of two spelling variants (kvinna and qvinna). In Table 1, we can see a summary

of the plots that are shown in Section 3.1. The terms are ordered on the basis of increasing frequency. An interesting behavior is that while the average Jaccard similarity increases by one order of magnitude between 'telephone' and 'music', the normalized frequencies have a similar increase between 'music' and 'travel'. With respect to correlation, 'happy' seems to be a trend breaker with a low correlation corresponding to a low instead of a high frequency. [2]

3.1 Jaccard similarities

In this section, we provide plots for each word in our study with the exception of 'rhetoric' where we have too little data to create yearly word embeddings. Each plot representing a word w can be read like this: The filled line is the smoothened Jaccard similarity, the dotted line is the normalized frequency. The values of the Jaccard similarity can be found on the left y-axis while the frequency values can be found on the right y-axis. In the title, we show the word and after it, the Spearman correlation value between the (non-smoothened) Jaccard similarities and the normalized frequencies.

Important to note when studying the plots is that a zero Jaccard similarity cannot be used to determine whether a word has a vector or not; for example,

[2] For a full answer to why these behaviors differ, many more words must be included in our study. This is left for future work.

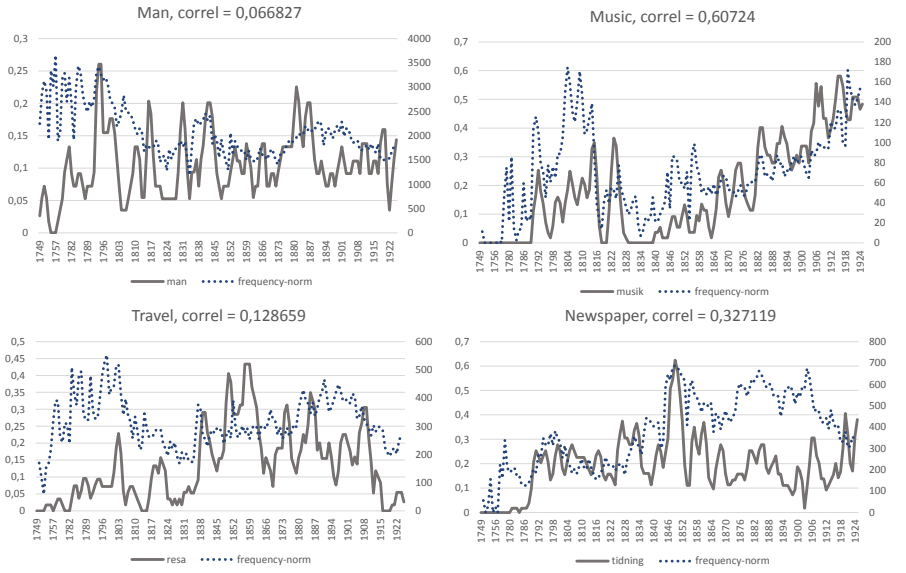

Fig. 1. Smoothened Jaccard similarities plotted against the normalized frequency of each word. The Jaccard similarity is on the primary axis (on the left side) and the normalized frequencies on the secondary axis.

the Jaccard similarity of ‘woman’ did not go above 0 until 1887 although the first vector appeared in 1859.

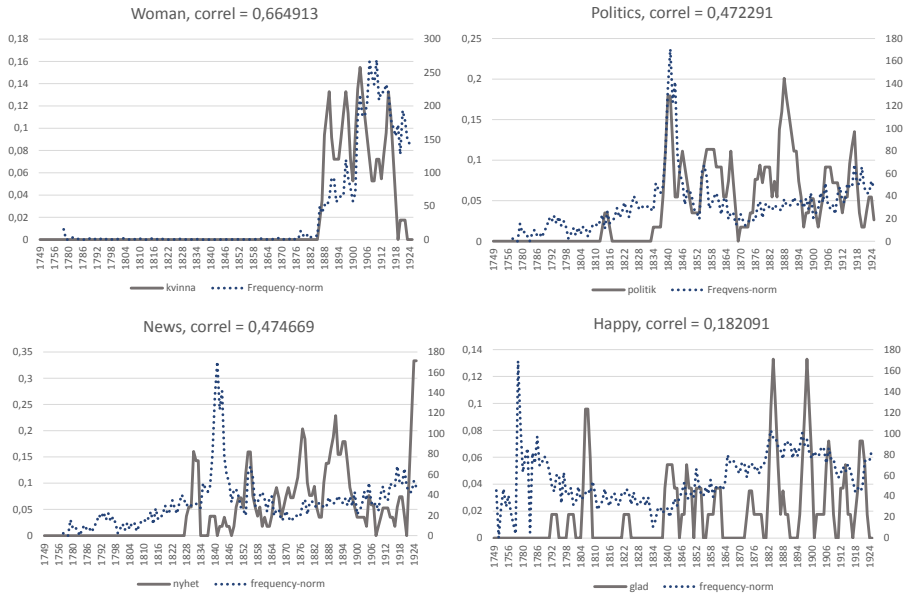

Fig. 2. Smoothened Jaccard similarities plotted against the normalized frequency of each word. The Jaccard similarity is on the primary axis (on the left-hand side) and the normalized frequencies on the secondary axis.

The plots in Figure 1 correspond to the most frequent words. Out of the four words, ‘music’ is the word that stands out with a high correlation. It is however the word with the lowest average frequency (see Table 1). For the word ‘newspaper’ we find that the frequency increases after the 1870s and the correlation between the two graphs is 0.6 for the years 1790-1877 as compared to 0.19 for 1878-1925. For the lower frequency words, ‘woman’, ‘politics’ (with the exception of the events in the 1880s), and ‘news’ (all in Figure 2) as well as ‘telephone’ and ‘telegraph’ (in Figure 3) , we see a high correlation between the Jaccard similarities and the frequencies. It seems that the more frequent the term, the lower the correlation with the Jaccard similarities. Reasonably, after a certain amount of data has been gathered, the embeddings become less dependent on the volume. And the lower the amount of data, the less stable the vectors.

Among the top frequent words, we find reasonably high Jaccard similarities each year, indicating fairly stable vectors. None-the-less, on average only 10-20 percent of all words are stable for each year.

The word ‘music’ has an interesting appearance, the correlation is high and both graphs show higher results before the 1830s, have a drop and then increase

slowly again. More investigation and close reading is necessary to determine what is happening in the 1830s and 1840s.

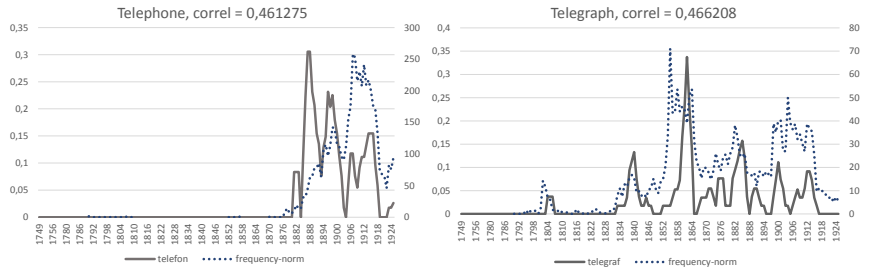

Fig. 3. Smoothened Jaccard similarities plotted against the normalized frequency of each word. The Jaccard similarity is on the primary axis (on the left-hand side) and the normalized frequencies on the secondary axis.

3.2 Tables for the top 10 similar words

In this section, we will present tables for each word and the top words that are most similar for a given year (we include as many of the top 10 words as will fit on the page; if all 10 words seem equally important, we will reduce the font size). Most years are not represented due to space constraints, but the first year is always present. For example, kvinna 'woman' had a W2V vector in 1859 and hence the top words for that year are included. The years chosen for each word will include the years with the highest Jaccard similarities, but also years that contain interesting words, so years will be different for each word. In each table we will include the top 10 words for the vector for the word when trained on the entire Kubhist corpus at once, called *all*.

year	words
1859	n&gt iinnu lz lllu &ltle lilliz &ltli&gt ll &lt&gt ssn
1860	-ne folt förbiflytande näpen gigg soin lwcnnc ätdon mellau nntcr
1877	påtagligen öfvervunnen ehuruväl alldaglig inspiration förstådda fördomsfri
1888	qvinna yvoffs flicka austins ung svägerska människa vuxen änka välmående
1903	flicka barn ung excentrisk sjuk våldta oregelbundet ansedd förföra tärd
1906	ung flicka halahult hjälplös orkeslös luggsliten rösträtt ädling röstresurser otukt
1910	människa rösträtt nyck ung nonchalant dödsfiende hederskänsla lika dem föibi
1911	kullkasta rösträtt samtid tapper karaktärsfast armod ung skicklig dryckenskap
1912	valbarhet valrätt rösträtt självförsörjande sexuell okunnig högerparti politisk radikal vänsterparti
1925	lik ung såra rädda föraktfullt roddbåt dragé hennes allvarligt medtagen
all	ung person flicka barn var två endast dem vuxen fullvuxen

Table 2. Table for kvinna 'woman'

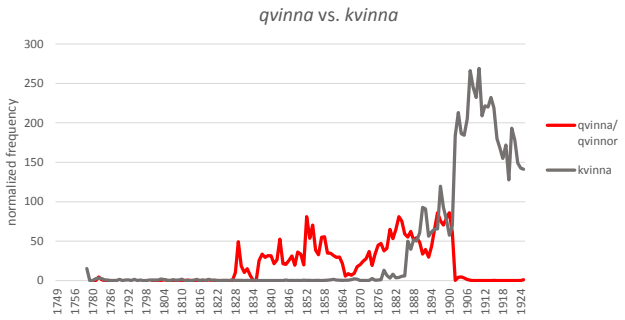

Fig. 4. Normalized frequency of the lemma kvinna and the words 'qvinna' and 'qvinnor', a previous spelling. We find that while both co-existed during a period, the 'qv' spelling was preferred before the 'kv' spelling took over.

Prior to 1859, a century in to our corpus, we cannot find any vectors for the word kvinna 'woman', which finds its explanation in the frequency of the word prior to 1859 being very low and mostly accidental (by means of spelling errors). This is due to spelling variations, where qvinna was the commonly used spelling. Figure 4 shows the mostly complementary frequencies of the lemma kvinna and the words qvinna and qvinnor. Before then, spellings like qwinna and kona were used. None-the-less, it seems women were mentioned more frequently toward the mid-end of the 19th century and the example shows the need for detecting language changes (spelling changes as well as sense changes) when analyzing historical texts.

When it comes to the top 10 words for kvinna 'woman', we find that the first few provide little reasonable content; the words are noisy with spelling errors. In 1888, the words are mostly descriptive of different kinds of women; 'young', 'sister in law', 'grown', 'human' and 'widow'. We find the first and only occurrence of våldta 'to rape' in 1903 (for women), together with words like 'girl', 'kid', 'young' and 'seduce'. The word 'to rape' is most likely lemmatized from 'raped', and the same goes for 'seduced'. In 1906, the word rösträtt 'right to vote' shows up among the top 10 words for 'woman', 13 years before women were allowed to vote in Sweden. Around that time, we see a strong increase in the frequency of woman, hence, they are more present in the newspapers. To complement, we have the top 10 words for the qvinna spelling in Table 3. The first vector appears in 1828 with words that have little to do with women. In 1837, we have a description of women with 'vale', 'beauty', 'naked' and 'abuse' which is most likely a lemma of 'abused'. In 1850, women are described with words that relate to their offspring, 'still born', 'twin', 'boy' and 'girl' while a year later we are back to a reasonably positive description, with 'lovable', 'loved', 'lover', and 'kissed'.

For the word politik 'politics' shown in Table 4, we find an interesting behavior around the end of the 1840s and 1880s. In 1838-1839, the Swedish historian and riksdagsman Erik Gustaf Geijer moved from conservatism to liberalism and

joined the movement for the common right to vote[3]. This might be the first spike that we see in frequency during this time. The second spike is likely due to a newspaper called politik 'politics' as seen from the quote:
i den i köpenhamn utkommande tidning " politik " '*in the Copenhagen-published newspaper "politics"*'.

For the word telefon 'telephone' we find high values of Jaccard similarity around the 1880s (in a period with a lower normalized frequency), which seems to be due to a Mr Håkan Bengtton (possibly Bengtson) who was a publisher for Göteborgs handelstidning, the Gothenburg trade paper. With low amounts of data, these kinds of peculiarities are seen more often. Telefonf is short for telefonförbindelse which was another way of saying 'telephone number', typically though in this format: *telefon : allm . telefonf . 519* .

year	words
1828	sång hufvud förgäfves inskränkta nedslagenhet öfverhus volontär olyckligtvis
1837	slöja späd hydda skönhet, skägg obändig naken förföra turban misshandla
1850	dödfödda dödfödd tvilling kön gosse äkta hicka lefvande promenerande flickebarn
1851	älskvärd kyssa tusenskälm älskad älskare uppfostrad jollra värdinna qvinlig körsven
1872	qvinlig ensamhet förtrollande tjusande motvilja svartsjuka drömmande vink skönhet
all	varelse egensinnig oerfaren hjärtlös ljushårig slafvinna gråhårig tillbedjare dygdig världsdam

Table 3. Table for qvinna/qvinnor 'woman'

year	words
1779	tilsammans fastställa runa biträdd gibraltar medborgerlig skatte ärva sand intagande
1838	intervention tadla åsigt anda parlamentarisk flerfaldigt politisk kraftig afgjord sansad
1842	opposition politisk konstitutionell tadla liberal handla mening grundsats söndring
1888	officiös press tysklands trontal bulgarien europeisk novoje organ bulgarisk rysslands
1891	socialdemokrati press politisk ståndpunkt statsman dementi frisinnad parlamentarisk makt
1925	näring trygghet kamp arbetarrörelse konservativ nationell strävan europa neutralitet
all	socialdemokrati utrikespolitik demokratisk försvarspolitik demokrati politisk parlamentarisk taktik

Table 4. Table for politik 'politics'

year	words
1877	karavan uppfinnare sirius springs konstruera walbenström uppfinning förlösning bell
1886	allm telefonf håkan aum rad handelstidning haka ansvarig bengtaon bengtton
1887	telefonf allm kungsportsavenyen hskan telegrafstation håkan handelstidning bengtton pen telegraf
1897	saltsjöbad vördsam allm no nygatan rikstelefon proviantgatan hansgatan götaborg
1924	åhlén tel almqvist nyman holm sthlm ss ken £ fröberg
1925	sigv pappershandel k ore larssons tomat purjolök rödlök lax sik
all	tel telef riks kikat rikst thunbarg bikst larmtorget rikstel storgatan

Table 5. Table for telefon 'telephone'

[3] https://www.sydsvenskan.se/2014-03-14/las-utdrag-ur-per-t-ohlssons-nya-bok-svensk-politik, from Per T Ohlssons new book, Svensk politik 'Swedish Politics'.

4 State of the Art

The first methods for automatic word sense change detection were based on context vectors; they investigated semantic density (Sagi et al. [6]) and utilized mutual information scores (Gulordava and Baroni [7]) to identify semantic change over time. Both methods detect signals of change but neither aligns senses over time or determines what has changed.

Topic-based models (where topics are interpreted as senses) have been used to detect novel senses in one collection compared to another by identifying new topics in the later corpus (Cook et al. [8]; Lau et al. [9]), or to cluster topics over time (Wijaya and Yeniterzi [10]). A dynamic topic model that builds topics with respect to information from the previous time point is proposed by Frermann and Lapata [11] and again sense novelty is evaluated. With the exception of Wijaya et al. who partition topics, no alignment is made between topics to allow following diachronic progression of a sense.

Graph-based models are utilized by Mitra et al. [12,13] and Tahmasebi [14] and aim to reveal complex relations between a word's senses by (a) modeling senses per se using WSI; and (b) aligning senses over time.

The largest body of work has been done using word embeddings of different kinds in recent years (Basile et al. [15]; Kim et al. [16]; Zhang et al. [17]).Embeddings are trained on different time-sliced corpora and compared over time. Kulkarni et al. [18] project words onto their frequency, POS and word embeddings and propose a model for detecting statistically significant changes between time periods on those projections. Hamilton et al. [19] investigate both similarity between a priori known pairs of words, and between a word's own vectors over time to detect change. [15,19,18] all propose different methods for projecting vectors from different time periods onto the same space to allow comparison. These methods can find changes in the dominant sense of a word but cannot differentiate between senses or allow some senses to stay stable while others change. The advantage of word embeddings over graph-based models, for example, is the inherent semantic similarity measure, where otherwise resources like WordNet are often used. We believe that the future lies in a combined approach, using embeddings (possibly multi-sense embeddings [20,21,22]) and sense-differentiated techniques.

5 Conclusions and Future Work

In this paper, we performed a study on (neural) word embeddings for a Swedish historical newspaper corpus, Kubhist. Our aim was to assess the quality of the Word2Vec model when the volume and quality of the text is limited, as is the case for most languages for historical contexts, English being the exception. Our timespan was 1749-1925, with the majority of the content being placed in the period 1850-1900. We investigated the stability, and through that, the quality of the resulting vector space for a set of 11 words. As a measure of stability, we use the word overlap between the top 10 most similar words for adjacent years. We

see a clear relation between the frequency of a word and the overlap from one year to another. The higher the frequency of a word, the higher the stability for the vectors. Conversely, the lower the frequency, the less stability we have.

None-the-less, even the highly stable words, 'music', 'travel', 'newspaper' and 'man' only have an average of 0.11-0.19 overlap (Jaccard similarity). This means that even the most stable words do not share many words in common from one year to another. This gives us reason to believe that the vector space produced by Word2Vec cannot be directly used for word sense change detection, in particular not if sense-differentiated embeddings are intended where the textual evidence for each word must be further divided into senses, thus decreasing the amount of available text for each vector.

Our findings are in line with those of [23] that point to the randomness that affects the outcome of embeddings like Word2Vec, both for the initialization as well as the order in which the examples are seen for training and of [24] that point to over fitting when there is too little data. For the Kubhist data, there are only five 10-year periods, between 1850-1890, with over 100 million tokens, thus limiting the possibility of finding changes in stable vectors corresponding to true word sense change.

One peculiarity that we notice is the spelling errors that are present in the top 10 word lists. This indicates that one future direction is the correction of spelling errors to increase the quality and volume of the text. Our current work aims to investigate a newly digitized version of Kubhist (which we have been promised in the near future by the Royal Library) to distinguish the role of OCR errors from spelling variations and measure the improvement when correcting for both, making use of embeddings based on Singular Value Decomposition [19], which is better equipped for handling historical texts and removing the randomness of Word2Vec.

Acknowledgments

This work has been funded in parts by the project "Towards a knowledge-based culturomics" supported by a framework grant (2012–2016; dnr 2012-5738) and by an infrastructure grant (SWE-CLARIN, 2014 – 2018; contract no. 821-2013-2003), both from the Swedish Research Council.

References

1. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
2. Språkbanken: The Kubhist Corpus. Department of Swedish, University of Gothenburg. `https://spraakbanken.gu.se/korp/?mode=kubhist`.
3. Tahmasebi, N.: W2V experiments on Kubhist. Språkbanken, Department of Swedish, University of Gothenburg. `http://hdl.handle.net/10794/word2vec-study-kubhist`.
4. Borin, L., Forsberg, M., Roxendal, J.: Korp – the corpus infrastructure of Språkbanken. LREC 2012 (2012) 474–478

5. Team, D.D.: Deeplearning4j: Open-source distributed deep learning for the jvm, apache software foundation license 2.0. http://deeplearning4j.org (2017)
6. Sagi, E., Kaufmann, S., Clark, B.: Semantic density analysis: comparing word meaning across time and phonetic space. GEMS '09, ACL (2009) 104–111
7. Gulordava, K., Baroni, M.: A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. GEMS '11, Association for Computational Linguistics (2011) 67–71
8. Cook, P., Lau, J.H., McCarthy, D., Baldwin, T.: Novel word-sense identification. In: Proceedings of COLING 2014, Dublin, Ireland (August 2014) 1624–1635
9. Lau, J.H., Cook, P., McCarthy, D., Newman, D., Baldwin, T.: Word sense induction for novel sense detection. In: EACL 2012. (2012) 591–601
10. Wijaya, D.T., Yeniterzi, R.: Understanding semantic change of words over centuries. In: Proc. of the international workshop on DETecting and Exploiting Cultural diversiTy on the social web. DETECT '11, ACM (2011) 35–40
11. Frermann, L., Lapata, M.: A bayesian model of diachronic meaning change. TACL **4** (2016) 31–45
12. Mitra, S., Mitra, R., Maity, S.K., Riedl, M., Biemann, C., Goyal, P., Mukherjee, A.: An automatic approach to identify word sense changes in text media across timescales. Natural Language Engineering **21**(05) (2015) 773–798
13. Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., Goyal, P.: 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, ACL 2014 USA. (2014) 1020–1029
14. Tahmasebi, N., Risse, T.: Finding individual word sense changes and their delay in appearance. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. (2017) 741–749
15. Basile, P., Caputo, A., Luisi, R., Semeraro, G.: Diachronic analysis of the italian language exploiting google ngram. In: Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016). (2016)
16. Kim, Y., Chiu, Y.I., Hanaki, K., Hegde, D., Petrov, S.: Temporal analysis of language through neural language models. In: Workshop on Language Technologies and Computational Social Science. (2014)
17. Zhang, Y., Jatowt, A., Tanaka, K.: Detecting evolution of concepts based on cause-effect relationships in online reviews. In: Proceedings of the 25th International Conference on World Wide Web, ACM (2016) 649–660
18. Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S.: Statistically significant detection of linguistic change. In: World Wide Web, ACM (2015) 625–635
19. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change
20. Trask, A., Michalak, P., Liu, J.: sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings. CoRR **abs/1511.06388** (2015)
21. Li, J., Jurafsky, D.: Do multi-sense embeddings improve natural language understanding? In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, ACL (2015) 1722–1732
22. Pelevina, M., Arefyev, N., Biemann, C., Panchenko, A.: Making sense of word embeddings. In: Proceedings of the 1st Workshop on Representation Learning for NLP. (2016) 174–183
23. Hellrich, J., Hahn, U.: Bad company - neighborhoods in neural embedding spaces considered harmful. In: COLING 2016. (2016) 2785–2796
24. Bamler, R., Mandt, S.: Dynamic word embeddings. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017. (2017) 380–389

Appendix – Top k word tables

year	words
1802	nyfiken ytterlig fransoscrne segla pästä dey afträdande toussaint befara iudarne
1823	telegraf-depescher vpanien lissabon bulletin corfu kapitulera rapportera ankommet
1862	hamburg kursnotering eonsols notera tclegramm london vexelkontor telegramm
1863	hamburg kursnotering telegramm consols paris eonsols börsföreningen london notera
1884	telegrafering direkt posto morning post boende persontåg mrd ©efle frånboende
1916	trådlös texas telegrafisk arkangelsk mäntyluoto graecia spanien lots kirkwall avsända
all	trådlös telefonförbindelse förbindelse linie dominion-liniens snabbgående cymrlk trådlos

Table 6. Table for telegraf 'telegraph'

year	words
1771	passera ankomma wärt frukta compagniet nödsakad besynnerlig corsica indra
1782	fång himmel hydda frögd nåd grymt hyf qval plåga gud
1807	dina purpur hjerta ditt smärta opp kärlek blick själ flit
1883	förtjust tjusande sorgsen hänryckt förälskad silfverklar herrlig förlägen godlynt
1886	obeskrifligt snyfta förlägen tröstande orolig snällt vemodigt sucka kärleksfull
1884	förtjust bedröfvad hungrig munter fröjd retligt ledsen herrlig förnöjd lycklig
1898	retlig blyg förskräckligt hungrig generad tankfullt gladt vidskeplig förtjust
all	gladt skön munter stolt idel förtjust fröjd glädje blid fager

Table 7. Table for glad 'happy'. In 1886, no words overlap and the words are counterintuitive; snyfta 'sob', orolig 'worried' and förlägen 'embarrassed'.

year	words
1750	högst fara kraft tyckas fattas sida igen kanna naturlig förundra
1847	tidningar införd inrikes posloch post post-och inriket postoch post-ici ost-
1848	tidningar inrikes införd ost- post-och postoch timing imikes posloch post
1902	boråt wermlands-tidningen illustreradt lindesberg sundsvalls-posten kåsör vpsala
1903	posten bohuslänning vestergstlands västergötlands ipsala annonsblad falu-kuriren
all	skriva dalpil correapondenten införd stockholms-tidningen for ö-tg dala-bladet nummer spalt

Table 8. Table for tidning 'newspapers'

year	words
1749	eller för med inkomne pä föda wid och död
1794	nia men han jag göra fiende skola näl förmodcligen sannolikhet
1877	han men öfverdrifvet förmätet märkvärdigt charlatan hon blodsutgjutelse tvifvelaktig obrottsligt
1918	men vi emellertid aldrig nog just det något kanske snart
1919	något gå vi men alltför längre nog någon ju kunna
all	nia men han vi kunna de även da dessa så

Table 9. Table for man 'man': Maximum Jaccard similarity is 0.43 for 1794, where mostly pronouns are overlapping. Fiende 'enemy' is the only content word that the two adjacent words have in common.