On the Uses of Word Sense Change for Research in the Digital Humanities

Nina Tahmasebi[1] and Thomas Risse[2]

[1] Språkbanken, University of Gothenburg, Sweden
nina.tahmasebi@gu.se
[2] Leibniz Universität Hannover, L3S Research Center, Germany
risse@L3S.de

Abstract. With advances in technology and culture, our language changes. We invent new words, add or change meanings of existing words and change names of existing things. Unfortunately, our language does not carry a memory; words, expressions and meanings used in the past are forgotten over time. When searching and interpreting content from archives, language changes pose a great challenge. In this paper, we present results of automatic word sense change detection and show the utility for archive users as well as digital humanities' research. Our method is able to capture changes that relate to the usage and culture of a word that cannot easily be found using dictionaries or other resources.

1 Introduction

When interpreting the content of historical documents, knowledge of changed word senses play an important role. Without knowing that the meaning of a word has changed we might falsely place a more current meaning on the word and thus interpret the text wrongly. As an example, the phrase *an awesome concert* should be interpreted as a positive phrase today. However, *an awesome leader* in a text written some hundred years ago, should be interpreted as a negative phrase, i.e., one to fear. The interpretation depends on the time of writing and not on the context terms and is thus not a pure disambiguation problem. Instead, we consider this as manifestation of **word sense change**.

The emergence of large digital and historical archives gives us a chance to learn these changes and to utilize them for research, both in linguistic research and in the digital humanities. It also gives us the possibility to feed our results back to the archives for better search and interpretation of results, thus opening them up for the public. Researchers can follow a word over time, query for specific kinds of change or mine for events that co-occur with language changes.

In this paper, we present and discuss results of automatic word sense change detection utilizing induced word senses. In Tahmasebi et al. [20] the induced word senses were evaluated on historical data and shown to provide good quality sense approximation. In Tahmasebi [21] we present the details of the word sense change detection algorithm. In this paper, we focus on analyzing and interpreting the results of word sense change.

We measure the time between an expected change in word sense and the corresponding found change to investigate not only *if* but *when* changes can be found and with which time delay. The delay aspect is of particular interest for linguists and concept historians. Why is there a time delay and how does it differ between regions, media and time? There is evidence that our language changes quicker in social media [8], can we see this also in modern traditional media? We believe that by capturing cultural changes in addition to sense changes, our results can be of importance for the digital humanities and social sciences.

2 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. [19]) 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 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. [2]; Lau et al. [12]), or to cluster topics over time (Wijaya and Yeniterzi [25]). A dynamic topic model that builds topics with respect to information from the previous time point is proposed by Frermann and Lapata [6] and again sense novelty is evaluated. Topics are not a 1-1 correspondence to word senses (Wang and McCallum [24]) and hence new induction methods aim at inferring sense and topic information jointly (Wang et al [23]). With the exception of Wijaya et al. that 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. [14,15] and Tahmasebi [21] 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 models allow us to identify individual senses at different periods in time and Tahmasebi also groups senses into linguistically related concepts.

The largest body of work is done using word embeddings of different kind in the last years (Basile et al. [1]; Kim et al. [10]; Zhang et al. [26]).Embeddings are trained on different time-sliced corpora and compared over time. Kulkarni et al. [11] 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. [9] investigate both similarity between a priori known pairs of words, and between a word's own vectors over time to detect change. [1,9,11] all propose different methods for projecting vectors from different time periods onto the same space to allow comparison.

Methods for detecting change based on word embeddings do not allow us to recover the senses that have changed and therefore, no way of detecting *what* changed. Most methods show the most similar terms to the changing word as a method to illustrate what happens. However, the most similar terms will only represent the dominant sense and not reflect changes among the other senses or capture stable parts of a word. The advantage of word embeddings over *e.g.* graph-based models is the inherent semantic similarity measure where otherwise often resources like WordNet [13] are used. In addition, compositionality methods can be used to find labels to help users better understand the results.

Due to a lack of proper evaluation methods and datasets, all presented papers have performed different, non-comparable evaluations. Most previous work have opted to pre-determine a set of words for further evaluation, both positive and negative examples of word sense change, rather than to evaluate the top terms outputted by the system thus needing evaluation for each new set of parameters.

3 Methodology

As a basis for our analysis we consider automatically induced word sense clusters. Each cluster represents a distinct time period and consists of a set of nouns and noun phrases of length two, i.e., **terms**. These clusters are approximations of word senses and to some extent capture also contexts. Throughout the paper we use **word senses** and **clusters** interchangeably. A **concept** consists of senses that are related (i.e., polysemous) following Cooper [3].

To model word sense change, we should allow each sense to change individually; worst case, this results in a graph where, for each time period $t \in T$ and a maximum number of sense S, we have in the order of S^T edges representing sense similarity. Even for a small number of time periods, this graph becomes infeasible to evaluate and investigate. Therefore, we reduce this complexity by first considering coherent senses over time (units) and then following the units over time. Units that are related are placed in a *path*. A unit can contain an arbitrary number of clusters, so to get a good representation of a unit, we create a centroid called a *unit representative*. We measure **similarity between units** as similarity between the unit representatives.

Individual *senses* s_w for a word w at one point in time are captured by *clusters*. A unit $u(w)$ captures a coherent sense s_w over a period of time and allows some change within s_w, e.g., broadening and narrowing. A path corresponds to a concept by grouping all units that are related (polysemous).

Our methodology consists of three steps. Firstly, deriving word sense clusters. Secondly, finding coherent senses by merging clusters into units and representing these with their unit representatives. Thirdly, relating units into paths by comparing unit representatives.

We find the word senses using an unsupervised word sense induction algorithm called *curvature clustering* (Dorow et al.[5]). The algorithm calculates clustering coefficient in a co-occurrence graph built with nouns and noun phrases that appear in the text separated with *and*, *or* and *commas*. Nodes with low clustering coefficient are removed and the graph falls apart into clusters that represent word senses. These clusters were shown to have 85% precision [20]. To the best of our knowledge, the curvature clustering method is the only induction method that has been properly evaluated on historical texts.

An Example For the details of the algorithm, we refer to Tahmasebi [21] and instead give an example to illustrate the workings. We start with three time points t_1, t_2, t_3 and unit sets $U_{t_1}(w) = \{u_1\}$, $U_{t_2}(w) = \{u_2, u_3\}$ and $U_{t_3}(w) = \{u_4, u_5\}$ for the target word *tape*. In this first iteration, each unit represents one cluster.
$u_1 = \{$*stereo, cassette, tape, record, radio*$\}$,
$u_2 = \{$*pin, thread, tape, silk, chair, cotton*$\}$,
$u_3 = \{$*video, cassette, tape, record*$\}$,
$u_4 = \{$*tape, sparkplug cable, wire, clip*$\}$,
$u_5 = \{$*television, record, tape, video, book, film, magazine, video industry*$\}$.

In the first step, similarity between pairs (u_1,u_2) and (u_1,u_3) is measured. Pairs are ranked according to the highest similarity and the pair with the highest similarity is merged. In this case, u_1 and u_3 are merged into $u' = \{u_1, u_3\}$ because u_3 is an almost subset of u_1. The unit representative consists of the terms $\{$*cassette, tape, record*$\}$. The pair (u_1,u_2) is removed because u_1 is already merged with one unit from $U_{t_2}(w)$.

The resulting merged set is $U_{[t_1,t_2]}(w) = \{\{u_1, u_3\} = u', u_2\}$. At time t_3, unit u_4 and u_5 are compared to the two units in $U_{[t_1,t_2]}(w)$. u_5 is merged with u' resulting in $u'' = \{u_1, u_3, u_5\}$. u_4 remains a single unit and is placed in $U_{[t_1,t_3]}(w)$ without being merged. When we merge two units, we add up all their clusters and build a new representative. When unit u_5 is merged with $u' = \{u_1, u_3\}$ we consider this to be a broadening because the single unit u_5 has a broader sense than the merged unit u'. The resulting unit set consists of $U_{[t_1,t_3]}(w)$=$\{\{u_1, u_3, u_5\}$=$u'', u_2, u_4\}$.

As a final step, to create paths, we measure similarity between the pairs (u'',u_2) and (u'',u_4). In this example, no units are related into paths which tells us that there are three different concepts for *tape*, one regarding *sewing tape*, one regarding *scotch tape* and one regarding *musical tape* which later includes also the *video tape*, matching well the main senses of tape but also capturing *sewing tape*, a sense less common today (OED [17]).

4 Experiments

The aim of our experiments is to find the quality and degree (i.e., recall) to which word sense change can be found using our word sense change detection and to investigate the utility of the results for research communities outside of linguistics. There exist no standard datasets or automatic evaluation metrics for word sense change. In addition, evaluation is a hard task because the outcome is specific to the collection and inherent location in mind; *when was a term used for the first time in the collection with the correct corresponding sense*? Therefore, in our experiments, we opt for a simplified, manual evaluation. We evaluate the found change for each term against the main changes of the term according to a set of knowledge sources and do not take completeness into account, *e.g.* by ignoring fine-grained sense differentiations.

We use *The Times Archive*, a large sample of modern English spanning 1785 – 1985. The collection is OCRed and corrected for OCR errors using the OCR Key method (Tahmasebi et al. [20]). We append the *New York Times Annotated Corpus*, a modern collection spanning 1987 – 2007, and disregard the annotations to treat both corpora the same. In total, the corpora span 222 years.

4.1 Testset

As a testset, we manually chose a set of 23 terms which we know have experienced word sense change during the past centuries. The main changes for each term were found using Wikipedia, dictionary.com and the Oxford English Dictionary, see extract in Table 1, and the automatically found changes were compared against the manually found counterpart. In addition, we considered the words *automobile, bitch, camera, car, cinema, computer, internet, mail, memory, phone, racism, record, train, travel.* We consider major changes in usage as well as changes to sense. In cases where multiple (fine-grained) senses were available, we opted to accept the widest sense. *E.g.* for the term *rock* we consider a *music* sense without any distinction between different types of rock music, because our dataset is unlikely to have fine-grained sense differentiations. If a clear time point cannot be pinpointed, we choose the earliest possible. For comparison purposes we also chose a set of 11 terms (*deer, export, mirror, symptom, horse, ship, paper, newspaper, bank, founder, music*) that have experienced minimal change during the investigated period, i.e., **stable terms**. The full testset can be found in [16].

We consider individual senses and their changes as separate events, *e.g.* an added sense and later a changed sense are two separate events. We have **35 change events** and **26 non-change events**. The change category consist of evolved senses (*e.g.* broadening and narrowing) and novel senses (related, i.e., polysemous senses and unrelated, i.e., homonymic sense).

The existing senses are also split into two categories, *existing -stable* (senses that belong to words that do not change over the entire dataset) and *existing -evo* (stable senses of words that have changes to their other senses).

4.2 Evaluation

For each experiment, we measure **recall** as the proportion of expected change events that were found; and **average time delay** as the difference in time between the *expected*, according to our ground truth, and the *found* events.

Recall is straightforward and measures the portion of expected change found, according to our ground truth. The expected time of change is trickier; true expected time of change for a given term is the first time that it was used in the collection with the correct corresponding sense. We do not know this time and therefore we approximate it using two different time points. The first expected time point is the *time of definition* or *time of invention* of a term w,

[3] This can be found in [18] and corresponds to usage change rather than lexical change.

Table 1. Description of change for some terms used in the evaluation. WWI occurred during 1914–1918, WWII occurred during 1939–1945.

Term	**Year**	**Description**
tape	1960-1965	Common household use
aeroplane	1908	First modern aircraft design
	WWI	First test as weapon
	WWII	Large scale war weapon
rock	1950-1960	Birth of rock-and-roll music
gay	1985-1990	Recommended instead of *homosexual*
tank	1916	First tank in battle
cool	1964	Slang used for self-control
flight	WWI-WWII	First commercial flights non-war related
	after WWI	Commercial aviation grows rapidly
mouse	1965	The computer mouse was introduced
	1980-1985	Common usage with computers like Macintosh 128K
telephone	1839	First commercial use in Great Western Railway
	1893	28k subscribers in Sweden, highest density in the world.[3]
	1914	USA twice the phone density than any other country.

$t_{DI}(w)$, in a given dictionary or knowledge resource. However, that an invention has been made does not necessarily correspond to newspapers reporting on it frequently. *E.g.* the *computer* was invented in its modern form in the 1940s, but was not mentioned in newspapers often in the early 40's, most likely due to WWII. Therefore, as a second expected time point, we consider the *first cluster evidence*, $t_{CE}(w)$, indicating the first time the term appears in a cluster and hence can be used for tracking. If the term is present with the corresponding sense in the collection before the time of the first cluster evidence, it means that it has either been mentioned very few times, or that the clustering algorithm could not find it. This time point represents the first possible time point for the tracking, given the curvature clustering algorithm for extracting word sense clusters. The true expected time lies in the interval $[t_{DI}, t_{CE}]$. Finally, we have the time point when our method detects the change event, $t_{found}(w)$.

The time delay is $T_{DI}(w) = t_{found}(w) - t_{DI}(w)$ and $T_{CE}(w) = t_{found}(w) - t_{CE}(w)$. The average time delay is summed over all words, $AT_{DI} = \frac{\sum_{\forall w} T_{DI}(w)}{|w|}$ and $AT_{CE} = \frac{\sum_{\forall w} T_{CE}(w)}{|w|}$.

Experimental set-up We differentiate between change events, stable senses of changing words and stable senses of stable words. We provide an upper limit to our change detection (Upper) by considering only if the change event is present in our units, disregarding the relation to other senses. This provides a measure of how much can be found in our clusters and implicitly measures the quality of the induction algorithm for change detection. For our change detection, we

Table 2. Recall and time delay for all terms in the testset, where *BC* is the best case and *All* is the all class experiments. The *value* in *bold* represents delay time from first cluster evidence AT_{CE} and the second represents time of definition AT_{DI}.

	Recall		Avg. time delay	
	Upper	*All*	*Upper*	*All*
Evolved sense	0.91	0.71	**4.9** – 17.4	**12.0** – 21.2
Existing – evo	1.00	1.00	**11.7** – 59.0	**11.7** – 59.0
Existing – stable	1.00	1.00	**2.7** – 20.5	**2.7** – 20.5
Average excl. stable	0.94	0.80	**7.1** – 30.7	**11.9** – 35.4
Total average	0.95	0.84	**6.3** – 28.7	**9.9** – 32.1

expect the evolved senses to appear inside a unit, the polysemic senses should be found within an existing path to illustrate the relatedness to other senses, and the homonym senses should be in their own path to show the lack of relatedness.

5 Experimental Results

We will present the experimental results on recall followed by average time delay.

5.1 Recall

Table 2 shows the recall of our experiment. Our upper bound shows that we are able to find 95% of all changes and stable senses among our clusters, giving us an upper bound on the recall of 95%. The only senses that are not found are the first senses for *Internet* and *computer*, and *bitch* in its offensive sense, most likely because of few mentions in the dataset.

For the change events, we are able to find 71% of them in the way we expect in relation to the other senses. The ones that are missing are the polysemous novel senses. By looking at examples from this class, it is obvious that the linguistic definition is very hard to detect automatically. *E.g.* the term *memory* in a digital sense is related to *human memory*, but rarely used in similar context. *mouse* used in a *computer mouse* sense has no words in common with the *animal* sense, *train* as a *mechanic train* with a locomotive differs largely from a *train of people or vehicles* (*e.g.* funeral train) and the *musical tape* is related to the *sewing tape* because of the shape but share no common words. Therefore, our method cannot place them in the correct path but chooses to place them in their own path. Excluding the polysemic senses, our recall is 92% for the change events.

Table 3 shows units for the term *rock* corresponding to three paths. The first unit represents the stone senses and the last unit the Rock, paper, scissors game both in their own path. The remaining three units are placed in a path for the *music* sense, $u_2 \rightarrow u_3 \rightarrow u_4$. A future direction of investigation is to find why the first music sense appears first in 1979.

Table 3. Extract of units for *rock*. Units display some internal clusters and terms.

Year	**Cluster terms**
	Unit $\boldsymbol{u}_1$: 1951-2003 (Stone)
1951	rock, sand, mud, clay, rain, ward, stone
1987	gravel, rock, sand, asphalt
1998	gravel, rock, sand
2003	dirt, calcined, clay, rock, stone, sand, gravel, moy sand
	Unit $\boldsymbol{u}_2$: 1979-2006 (First music cluster)
1979	rock, jazz, marriage, advice bureau
1987	classical, soul, drug, rockabilly, sex, folk, funk, gospel
1995	jazz, reggae, rock, funk, rap, hard rock, punk
2006	chamber music, bluegrass, soul, blue, funk
	Unit $\boldsymbol{u}_3$: 1987-2003 (Modern music)
1987	rap, opera, calypso, drug, sex, drama
1995	grunge, punk, alternative rock, hiphop, blue, rock
2003	irish music, mexican, mixing rock, appalachian song, rock, hiphop
	Unit $\boldsymbol{u}_4$: 1988-2007 (Rock&Roll lifestyle)
1988	rock, roll, sex, african, drug
2001	fantasy of sex, sex, rock, drug, roll, capture
2006	guitarist, songwriter, freeassociates about religion, rock,
	Unit $\boldsymbol{u}_5$: 2000-20075 (Game)
2000	rock, paper, scissors

False positives Precision is not well understood w.r.t. word sense change detection when units can consist of 70-80 clusters and paths can contain hundreds of units. Instead, we analyze false positives by looking at the average number of change events per word. On average there are 3 paths per word and 5.3 units per path for change words and 13.3 for stable words. Among the changing words, we have an average of 2.2 change events and thus we would expect around 2 false positives (5.3 units mean 4 change events on average out of which we expect 2 to be correct). Among the stable words, all change events and thus different units are per definition wrong, that means on average 13.3 false positives. However, there are some words that stand out, *horse*, *bank* and *music* are very common words and have, in average, 47.5, 21.4 and 24.9 units per path when we would expect only one. For these we observe very long spanning units with 206, 197 and 204 years. Excluding these words, the average number of unit per path drops to 6.6 and represents 5 change events.

Though this is an approximation of the false positive rate, it does tell us that the number of elements to manually filter is limited and thus the results can be of great use for researchers and digital archive users. The true utility of the paths will be determined in future work with researchers from the digital humanities as well as normal users of digital archives.

5.2 Average Time Delay

Table 2 shows the average time delays for our experiments. Values marked in bold are delay times with respect to first cluster evidence, AT_{CE} and the second values are with respect to time of definition AT_{DI}. At best, we can find evidence in our units 7.1 years after the time the changes appear in our clusters and 30.7 years after being invented or defined in a dictionary. We consider the true time delay to be between 7.1 – 30.7 years. To appear in the paths as we expect, the time delay is slightly longer, between 11.9 – 35.4 years. If we split the time delays into the change categories, we have 16.1 – 20.9 year for the evolved senses, 5.8 – 27.8 for the polysemous senses and 1.6 – 19.8 years for the homonymic senses.

For **existing senses** we see something interesting; the existing senses for words that later have a change event have significantly longer average time delays compared to existing senses of stable terms, 11.7 compared to 2.7. One possible explanation is that words are less likely to change their meanings, if they are commonly used. The long time delays compared to definition is likely due to the choice of words in the stable category. The papers might not often discuss the *bitch* as a female dog, *train* as a train of people or the *car* as a wheeled, usually horse-drawn conveyance and hence we cannot detect these senses with our induction method, thus the longer time delays for stable senses of evolving words. On average, we find that excluding the existing senses of stable terms we have an average time delay of 7.1 – 30.7 years for any evidence to appear in a unit, 11.9 – 35.4 for our method to find the change in its expected form. Including existing senses, delay times decrease to 6.3 –28.7 and 9.9 – 32.1 respectively.

6 Discussions

Our experiments show that we are able to find much of the expected word sense changes as well as the stable senses. We depend on automatically induced word senses that are grouped into *units* to capture individual senses over time. Units are then grouped into *paths* that capture concepts for a term.

The utility of using a method that differentiates between senses of a word are plentiful. For example, the word *rock* has a stable sense of *stone* in our dataset and then, in the 20th century, adds a sense of *music style*. The *music* sense evolves with different kinds of music and adds a *rock-and-roll lifestyle* sense in the same path as the *music* sense, clearly showing that these senses are related.

Also among words that are considered the same meaning over time, we can find changes that reflect usage and culture. For example, the *telephone* was firstly mentioned in contexts that related to the entire community or to houses in general, *1882, hydraulic lift, electric light, telephone, lift*. Then, slowly, it became something that belonged inside each apartment, *1977, television set, freezer, telephone, refrigerator, cooker, washing machine* and then a tool for (mass) communication *1997, telephone, television, radio, newspaper*. The word *aeroplane* is firstly defined as a flying machine, *1908 airship, aeroplane, balloon, aeroplane construction*, then as a means of transportation *1914 plane, aeroplane,* *motor bicycle, motor lorry, car* and finally as a weapon of war *1917 piping, gun, aeroplane, shafting, tank, infantry.* The word *travel* had only senses related to a literature genre *1803 literature, science, art, travel, voyage* before we could see evidence in the early 20th century of actual travel *1906 full board, travel, best hotel.* It is important to note that our datasets represent different dialects, British (The Times) and American (New York Times) which could lead to changes that are due to dialectal differences rather than sense changes. Among our test set, we have only three words (*gay*, *phone* and *telephone*) where the expected change lies in the period up to 1985 (The Times) and the found changes is in the period after and hence bridges this dialectal gap. In addition, for the *All* experiment, the *computer* sense of *mouse* was found in 1995, the expected was in the 1960s and the first cluster evidence in 1985. For the remainder of the words, the expected and found changes lie in the same dataset and hence they do not suffer from risk of dialectal interference.

The results of word sense change can be used to help users of a digital archive to understand the content in the archive when the language has changed over time. Senses that have changed can be marked and examples can be presented to help interpret the older sense. Language changes will be an increasing problem as we store more social media content in our archives [8]. The advantages of automatically detecting sense changes from the archive directly rather than relying on an outside reference, *e.g.* a dictionary are also obvious; dictionaries are meant as references and do not model how people use the language. But the results of word sense change detection can also be useful for exploring an archive and the culture represented there; *E.g.* what was the updake of the *telegraph*?[4]. They can also be used for language teaching and learning [4].

There is a need for temporal sentiment analysis which can only be made reliably after having detected word sense change, to be able to differentiate between *awesome leaders* of different times but also to answer research questions like what the attitude towards *rhetoric* has been over time [22].

7 Conclusions and Future Work

In this paper we presented results for a word sense change detection method that relies on induced word senses as a basis for detecting word sense change. We present analysis of the results and show that these can have an impact for research also in the digital humanities, where the *when*, *how* and *why* of language change are important. We show that our method, in addition to finding word sense change, also finds cultural and usage change. Our method detects change in the correct form 11.9 years after the first cluster evidence and is the first work to report such time analysis. Given the 222 year timespan, we consider this delay to be a good starting point for future work and for analysis regarding differences between data sources, place of publication and time periods.

[4] `https://sweclarin.se/sites/sweclarin.se/files/videos/invigning_2016/Johan-Jarlbrink.mp4`

It remains future work to find the best way to preserve and utilize found change. Temporal indexing structures, information retrieval and presentation techniques as well as scalability issues are future directions for research in the field of automatic detection of word sense change. Preferably, digital archives should be stored with existing concurrent dictionaries and resources, and be word senses disambiguated to ensure long-term semantic access.

Acknowledgments

This work has been funded in parts by the project "Towards a knowledge-based culturomics" supported by a framework grant from the Swedish Research Council (2012–2016; dnr 2012-5738). This work is also in parts funded by the European Research Council under Alexandria (ERC 339233) and the European Community's H2020 Program under SoBigData (RIA 654024). We would like to thank Times Newspapers Limited for providing the archive of The Times for our research.

References

1. 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)
2. Cook, P., Lau, J.H., McCarthy, D., Baldwin, T.: Novel word-sense identification. In: Proceedings of COLING 2014. pp. 1624–1635. Dublin, Ireland (August 2014), `http://www.aclweb.org/anthology/C14-1154`
3. Cooper, M.C.: A Mathematical Model of Historical Semantics and the Grouping of Word Meanings into Concepts. Computational Linguistics 32(2), 227–248 (2005)
4. Dejica, D., Hansen, G., Sandrini, P., Para, I.: Language in the Digital Era. Challenges and Perspectives. De Gruyter (2016)
5. Dorow, B., Eckmann, J.p., Sergi, D.: Using curvature and markov clustering in graphs for lexical acquisition and word sense discrimination. In: Proceedings of the Workshop MEANING-2005 (2005)
6. Frermann, L., Lapata, M.: A bayesian model of diachronic meaning change. TACL 4, 31–45 (2016)
7. Gulordava, K., Baroni, M.: 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. pp. 67–71. GEMS '11, Association for Computational Linguistics (2011)
8. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Cultural shift or linguistic drift? comparing two computational measures of semantic change. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2016)
9. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. CoRR abs/1605.09096 (2016)
10. 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)

11. Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S.: Statistically significant detection of linguistic change. In: Proceedings of the 24th International Conference on World Wide Web. pp. 625–635. ACM (2015)
12. Lau, J.H., Cook, P., McCarthy, D., Newman, D., Baldwin, T.: Word sense induction for novel sense detection. In: EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 591–601 (2012), http://aclweb.org/anthology-new/E/E12/E12-1060.pdf
13. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38, 39–41 (1995)
14. 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), 773–798 (2015)
15. 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. pp. 1020–1029 (2014), http://aclweb.org/anthology/P/P14/P14-1096.pdf
16. Nina Tahmasebi, Thomas Risse: Word Sense Change Test Set. https://doi.org/10.5281/zenodo.495572 (2017)
17. OED, O.E.D.: http://www.oed.com/view/Entry/197656?rskey=8IY6gT$&$result=1$&$isAdvanced=false#eid (2017), [Online; accessed 2016-05-02]
18. Roslin Bennett, A.: The Telephone Systems of the Continent of Europe. Longmans, Green and CO. (1895), http://archive.org/stream/telephonesystems00bennrich#page/332/
19. Sagi, E., Kaufmann, S., Clark, B.: Semantic density analysis: comparing word meaning across time and phonetic space. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics. pp. 104–111. GEMS '09, ACL (2009), http://dl.acm.org/citation.cfm?id=1705415.1705429
20. Tahmasebi, N., Niklas, K., Zenz, G., Risse, T.: On the applicability of word sense discrimination on 201 years of modern english. International Journal on Digital Libraries 13(3-4), 135–153 (2013), http://dx.doi.org/10.1007/s00799-013-0105-8
21. Tahmasebi, N.N.: Models and Algorithms for Automatic Detection of Language Evolution. Ph.D. thesis, Gottfried Wilhelm Leibniz Universitt Hannover (2013), http://edok01.tib.uni-hannover.de/edoks/e01dh13/771705034.pdf
22. Viklund, J., Borin, L.: How can big data help us study rhetorical history? In: Clarin Annual Conf. (2016)
23. Wang, J., Bansal, M., Gimpel, K., Ziebart, B.D., Clement, T.Y.: A sense-topic model for word sense induction with unsupervised data enrichment. TACL 3, 59–71 (2015)
24. Wang, X., McCallum, A.: Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 424–433. KDD '06, ACM, USA (2006)
25. Wijaya, D.T., Yeniterzi, R.: Understanding semantic change of words over centuries. In: Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web. pp. 35–40. DETECT '11, ACM, New York, NY, USA (2011)
26. 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. pp. 649–660. ACM (2016)