A Framework for Analyzing Semantic Change of Words across Time

Adam Jatowt[1,2] and Kevin Duh[3]

[1]Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

[2]Japan Science and Technology Agency
4-1-8 Honcho, Kawaguchi-shi,
Saitama, 332-0012 Tokyo, Japan

[3]Nara Institute of Science and Technology
8916-5 Takayama, Ikoma
Nara 630-0192, Japan
kevinduh@is.naist.jp

ABSTRACT

Recently, large amounts of historical texts have been digitized and made accessible to the public. Thanks to this, for the first time, it became possible to analyze evolution of language through the use of automatic approaches. In this paper, we show the results of an exploratory analysis aiming to investigate methods for studying and visualizing changes in word meaning over time. In particular, we propose a framework for exploring semantic change at the lexical level, at the contrastive-pair level, and at the sentiment orientation level. We demonstrate several kinds of NLP approaches that altogether give users deeper understanding of word evolution. We use two diachronic corpora that are currently the largest available historical language corpora. Our results indicate that the task is feasible and satisfactory outcomes can be already achieved by using simple approaches.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing: *Linguistic processing*

Keywords

language change, word meaning evolution, historical linguistics, computational etymology

1. INTRODUCTION

Human language is subject to constant evolution driven by the need to reflect the ongoing changes in the world and to become a more efficient means of communication. Words acquire new meanings and cease to be used according to the old meanings. This is evident for anyone who wishes to read old texts, even, the ones created only few decades ago. Properly understanding the nature of changes in the meaning and usage of words is thus important for anyone who works with historical texts, such as librarians, historians or linguists.

Knowledge of word and language evolution is also a subject of interest to the general public, since language is a basic communication tool for expressing and sharing our thoughts. These are evidenced by the large number of websites devoted to etymology studies, and many books discussing word origins [1,3,4,10,12,14,16,24]. Moreover, understanding of language evolution can potentially have impact on Natural Language Processing (NLP). For example, it has been theorized that polysemy (a perennial challenge for NLP research) is simply a transitory phase of word evolution where a new sense is acquired and starts to compete with existing senses for "survival" in popular usage [2]. Last but not least, OCR algorithms could be improved when enhanced with automatically generated information on the scope of word meaning over time.

We thus advocate that language evolution and, in particular, word etymology is a worthy problem to study for specialists, general public, and NLP researchers alike. However, until now such studies were generally limited to relatively small scale and fragmentary investigations of individual words, phrases or grammatical patterns. This is because the analysis was based on laborious work consisting of locating the occurrences of features in old textual artifacts and then manually comparing their context and other characteristics. But we are now at an inflection point. Given the large amounts of digitized old books and texts available nowadays, we expect growing excitement of the potential for computational etymology.

Automatically capturing the evolution of word meaning across time and providing credible evidences of their explanations is however non-trivial. Given typically long time spans of analysis and the need for correct representation of each time unit, the underlying data must necessarily be huge as well as more or less uniformly distributed across time. This situation calls for rather simple and computationally light approaches. On the other hand, semantic transitions tend to be intricate and may not be detected or understood in-depth when employing only one-sided analysis.

We then propose a framework for exploring from multiple perspectives the lexical change, i.e. shifts in word meaning over

978-1-4799-5569-5/14/$31.00

time[1]. Using word representations from distributional semantics [11], we present algorithms to discover change from diachronic corpora at three different levels: single word (lexical) level, contrastive word-pair level, and sentiment orientation level. Together they provide the user an informed visualization of semantic evolution for any query word and form basis for inferring conclusions as for the history of the word. Our approach is exhaustive in the sense that we portray a word change in its intermediate stages instead of focusing only on the word's meaning at the time of its origin and at the current time. We demonstrate the effectiveness of our framework based on well-known examples of words that changed meaning over recent time. In order to obtain more comparative evidence we apply our methods on two different, large scale language corpora.

The remainder of this paper is structured as follows. Section 2 describes previous work related to the nascent field of computational etymology. Section 3 introduces our datasets and their preprocessing methods. The main part of the paper is Section 4, which is divided into several subsections, discussing our proposed algorithms for discovering each of the three aforementioned levels of semantic change. These subsections are augmented with case studies of interesting phenomena found by the algorithms. Finally, Sections 5 and 6 provide discussion, conclusions and the outline of future work.

2. RELATED WORK

2.1 Language Change

The language is in a constant flux of change. It is being adapted to the changing real world which it is supposed to describe. Another reason driving the change is the continuous process of increasing efficiency of expressions according to the least-effort principle proposed by George Kingsley Zipf [28]. In the context of linguistic communication this rule states that the need for the decrease of the required effort in conversation happens at both the speaker and listener sides.

The current state of the language should be considered as the cumulative outcome of communication occurring within previous generations. Over time new words were either coined or borrowed from other languages [14,24]. On the other hand many existing words were subject to semantic change to let speakers refer to new or changed concepts in the surrounding world. In this work, we are interested in the latter type of linguistic evolution.

Diachronic linguistics (aka *historical linguistics*) is an area within modern linguistics that concerns itself with the process of language development over time and, in particular, with questions why and how languages change and in which way these changes spread across spatio-temporal dimension [1]. It is often compared with *synchronic linguistics* which, in contrast, studies the state of the language at given time, commonly, at present time. Both subfields are in close relationship with each other.

Careful studies carried within diachronic linguistics resulted in the wealth of knowledge about the histories of words. However, the large sizes of vocabularies of many languages (in the range of tens of or even hundreds of thousands of words) and different meanings of individual words make this work far from complete. In general, due to the long neglect of semantics - the science of meaning - within the linguistics, still, there is no clear conception of exact principles governing the meaning change in language [1,24].

So far three general theories have been proposed to explain the reasons behind the actuation of linguistic phenomena: *functional*, *psycholinguistic* and *sociolinguistic* explanations. The first one assumes that language has natural tendency to "regulate itself" and thus becomes more regulated, symmetrical and simpler over time. One example is the observed avoidance of homonymic clash – a situation when two homonyms, words with same form but different meaning, exist at the same time. On the other hand, the *psycholinguistic explanation* associates the language change with the cognitive processes occurring in the brain of a speaker, which is affected by memory limitations or processing procedures. Finally, the *sociolinguistic theory* [12,14] explains the changes in the meaning of used words by the social circumstances and their variations over time. For example, social context and the formality of situations are well-known to affect the way in which speakers express their thoughts.

2.2 Computational Approaches

Recently, there has been growing interest in applications of language technology for cultural heritage and humanities. Michel et al. [18] introduced *culturonomics* – the study of cultural and historical phenomena based on large textual data. The authors demonstrated rising or falling frequencies of selected words that allow to reason about higher level cultural or abstract changes occurring in society. For example, they contrasted the popularity plots of word "men" vs. the one of "women" to provide evidence for the increasing social role and emancipation of women in recent decades. The objective of our work is fundamentally different as we seek to study the evolution of word meaning by investigating its context across time. We are also more interested in studies of the contextual neighborhood of words as seen from etymological perspective, rather than in their popularity measured by the differences in word frequency over time.

Previous work on computational approaches in the context of etymological studies has been very scarce. The Google Books N-gram Viewer [2] is useful only for observing counts of word appearance through time so it fails to elucidate how the word was used and when its meaning transitions occurred. The online interface to Time Corpus of American English[3] created by Brigham Young University [5] goes one step further, as it allows seeing examples of keywords in context (KWIC) at different time points. However, the amount and the detail level of the shown data are prohibitive for efficiently inferring broader conclusions about the word from the longitudinal perspective.

Odijk et al. [22] demonstrated interactive environment that visualizes information on volume and correlation of words and documents across time. Like [18] their focus is leaning more towards understanding historical and social aspects rather than the evolution of word semantics. Also, their proposal does not incorporate across-time sentiment analysis, comparative word analysis or different word representations that we utilize.

Other works do not directly assume the objective of explaining semantic evolution at the word level yet are to certain extent related to our research. Mihalcea and Nastase [19] proposed to identify epoch to which a given word belongs. They automatically classified a word to any of the three epochs in the past based on word's context. Gerow and Ahmad [8] reported trends in language which manifest at the token, part-of-speech and grammatical levels based on the corpus of NIPS conference papers, while Mair et al. [17] studied short term diachronic shifts in part-of speech frequencies.

Tahmasebi et al. [25] introduced NEER, an unsupervised method for the named entity evolution recognition that is independent of external knowledge sources. Their goal is to automatically find

[1] Other areas of interest in language evolution include syntactic change, sound change, areal effects/borrowing, and mathematical models of evolution. We focus on lexical change.

[2] https://books.google.com/ngrams

[3] http://corpus.byu.edu/time

different lexical names for the same given entity. For example, Hillary Clinton should be matched with Hillary Rodham and Pope Benedict XVI would correspond to Joseph Ratzinger as terms that mean the same thing at different points in time.

Jatowt and Tanaka [13] studied language change from the macroscopic viewpoint. They emphasized the rich-get-richer effect in word popularity over time which means that highly frequent words tend to remain frequent over time.

Regarding the sentiment analysis part, the concept of estimating positive or negative polarity of words has been around for some time [15,20,26]. However, the previous works focused mainly on synchronic analysis [15,26] while we apply it for the diachronic scenario. Mohammad [20] presented an emotion analyzer for studying sentiment in fairy tales and novels. He proved that fairy tales have a much wider range of emotion word densities than novels using the proposed emotion density measure inside texts. The entity's emotion associations are derived from their co-occurring words. The author also demonstrated the percentage of fear, joy and anger related words that appear over time in association with some selected words.

3. DATA AND PREPROCESSING

3.1 Datasets

We have used Google Books 5-gram dataset[4] (version 2009) which is the largest available historical corpus that provides a comprehensive capture of language use in the printed world. The dataset has been compiled over about 4% of ever published books [19]. It spans the time period from 1600 to 2009 and, in total, has the size of over 1TB of text data (about 0.3 trillion words). The rate of OCR errors is limited since n-grams that appear over 40 times across the corpus have been removed. The Google Books dataset nicely provides n-gram count information by year. For efficiency and for comparison purpose with the other datasets, we mapped the yearly granularity of the 5-gram data to the decade granularity and we used only the part ranging from 1800 to 2009 as it has the largest number of n-grams. The year-decade conversion was done by summing unique 5-grams over all the years in each decade. We could have alternatively opted for a more complex sliding window sum, but this straightforward decade-level view is advantageous for our objective, as it is intuitive and is also detailed enough to capture any long term changes while removing short term fluctuations that quickly fade away (e.g., ones driven by short-lived events).

Our study was supplemented by much smaller corpus: Corpus of Historical American English[5] (COHA) [5], which is a balanced diachronic corpus compiled over diverse document genres. COHA contains over 400M words. They were collected from about 107K documents which were published from 1810s to 2000s and are available in the form of n-grams. In this study we used 4-grams dataset. The documents used for COHA were carefully selected by maintaining the fixed ratio of different genres throughout different decades following the Library of Congress classification [6]. According to the COHA's creators, the corpus is 99.85% accurate, which means, on average, there is one error for about every 500-1000 words [5]. COHA dataset provides the frequency of each ngram for every decade.

We emphasize that both datasets are substantially different. Google N-gram dataset has been compiled using books digitized within the Google Books initiative. In contrast, COHA contains carefully selected prose texts and it is characterized by a relatively stable rate of different document genres across decades. However, many rare words are not present in COHA making it useful only for the analysis of relatively common terms. On the other hand, Google Books 5-gram dataset is more flexible in this regard as well as more reliable given its size. Thus in this paper we mainly base our analysis on Google Books 5-gram dataset and we occasionally verify the results by comparison with ones obtained from COHA.

The further preprocessing steps for both datasets involved converting words to lower cases and removing digits and other non-word tokens. We also discarded stop words based on the stop word list provided by Natural Language Processing Toolkit[7] (NLTK). The choice of stop words needs to be done carefully due to time passage[8]. The NLTK stop word list is convenient as it is minimal (127 words) and contains words that have the highest or very high across-decade frequencies within the study period we focus at.

3.2 Word Representation

Word representation lies at the heart of any lexical semantics work. We adopt here a distributional semantics view, where each word is represented by its usage context, that is, its neighboring words in the n-gram datasets. Distributional semantics is best portrayed by a well-known saying attributed to J. Firth [7]: "You shall know a word by the company it keeps". In our implementation we compute a vector representation $d_i[w]$ for each word w in each decade i. Intuitively, if the similarity between the same word's vector at decade i and decade j is low (i.e. $sim(d_i[w], d_j[w]) \rightarrow 0$), then a semantic change is likely to occur between these decades. This representation also enables us to compute changes in similarity between pairs of contrasting words w_1 and w_2, e.g., $sim(d_i[w_1], d_i[w_2])$. We propose to examine three different kinds of word representations. They are described below.

3.2.1 Normal Word Representation

The first one is called *normal representation* and is the conventional way of capturing distributional information without considering word order. For a given word w in a decade d, we collect all 5-grams containing the word in any position in this decade. Then we sum the counts of neighboring words. The word representation is a vector of size constrained by the number of unique words in the dataset. The weights in the vector are calculated as the counts of words co-occurring with the target word w in d_i divided by the total number of co-occurring words with w in that decade.

3.2.2 Positional Word Representation

The second representation, *positional representation*, captures not only the frequency of co-occurring words but also their relative positions in relation to the target word. The way to compute it is based on extension of the normal vector representation such that each word is associated with index that expresses the position of the context word with respect to the target word, which is always fixed to the index position 0. The relative positions of any term regarding the target word are then: -3, -2, -1, +1, +2, +3 for COHA and -4, -3, -2, -1, +1, +2, +3, +4 for Google Books 5-gram datasets. Figure 1 depicts the index values of a context word in a 5-gram. For computing the positional vector representation we calculate the count of each context word in its corresponding position. Thus the positional vector size is 8 times longer than the normal one for the Google Books 5-gram dataset and 6 times larger for the case of COHA dataset. Unlike the normal vector representation, the positional representation offers more fine-grained order

[4] https://books.google.com/ngrams/datasets

[5] http://corpus.byu.edu/coha/

[6] http://www.loc.gov/catdir/cpso/lcco/

[7] http://nltk.org

[8] Some words could be considered as stop words at certain time, yet as content words at another time

information, which may be useful in the absence of part-of-speech or syntax information. Note that syntactical annotations of the datasets are available now [9] and we hope to incorporate these in our word representations in future work.

```
w - - - a  +4
w - - a -  +3
w - a - -  +2
w a - - -  +1
a - - - w  -4
- a - - w  -3
- - a - w  -2
- - - a w  -1
```

Figure 1 Possible index positions (in red) of a context word *a* in 5-gram. We assume *w* is a target (query) word.

3.2.3 LSA Based Representation

The last vector representation, *LSA based representation*, is based on Latent Semantic Analysis [6]. The motivation for LSA is to gracefully handle rare words, in cases where the word counts may be too sparse for normal and positional representations. The procedure is analogous to LSA performed on document-term matrix, but with a slight modification since the ngram-term matrix in our case is too large to fit in memory. Following [23], we compute the term-term correlation matrix $\boldsymbol{C}$ where each term is a conjunction of word w and the decade i. Each element (y,z) of $\boldsymbol{C}$ is the number of times words y and z co-occur in the same n-gram in the same decade. Given a vocabulary size of $|V|$ and 20 decades, this is a matrix of dimension $20|V|$ x $20|V|$. Finally, we use sparse SVD to compute the k largest left eigenvectors of $\boldsymbol{C}$, and use them as the k-dimensional word representation $d_i[w]$ for each word-decade term. Note that it is important to perform the LSA jointly across time, or else $d_i[w]$ and $d_{i-1}[w]$ would not have any shared elements for computing similarity.

Our methods described in Section 4 work with any of the three word representations introduced above. In the current implementation, due to high computational cost, we computed the LSA based representation of a word only for the COHA dataset.

4. FRAMEWORK

In this section, we describe in detail our methods for discovering semantic change at three different levels. First, we capture semantic change of a single word through context comparison across decades (Section 4.1). Then, we extend this to find evolution of similarities between contrasting pairs of words (Section 4.2). Finally, we show how to discover changes in sentiment orientation (Section 4.3), which is one of the main ways in which words change. While each method is independent, they are coordinated by time, and, together, they offer an informed visualization of word evolution for the user to support etymological analytics.

We complement the discussion of the proposed methods with analysis results of example words whose semantic evolution is well-known. Most of the words shown in the case studies have been manually picked up from the *Oxford Dictionary of Word Origins* [4], while the rest were collected from the *Online Etymological Dictionary* [10]. We selected the words that underwent significant meaning changes within the period of our study. The results that we show have been obtained using Kyoto University Supercomputer[9] - a parallel processor (MPP) system connected through a high-speed network to 940 nodes with 32 cores and 64GB memory per node. It can deliver calculation performance of 300 teraflops and a memory capacity of 59TB.

[9] http://www.iimc.kyoto-u.ac.jp/en/services/comp/

4.1 Semantic Change of Single Word

For studying the semantic change of individual words across time we construct vector representation of word context in each decade as discussed in the previous section. Then we compare the context of the word in the last decade with the ones in the previous decades. Figure 2 shows the concept behind this comparison. We apply here cosine similarity calculation as it is well-known and commonly used.[10] The proposed visualization allows us to see how the meaning of a word evolved along with the time flow by observing the shape of the similarity curve before it reaches value of 1 at the point of the last decade. The curve with high and steep increase in the similarity characterizes a word which rapidly acquired the present set of meanings. On the other hand, a flat curve over time points indicates words with stable meaning and usage.

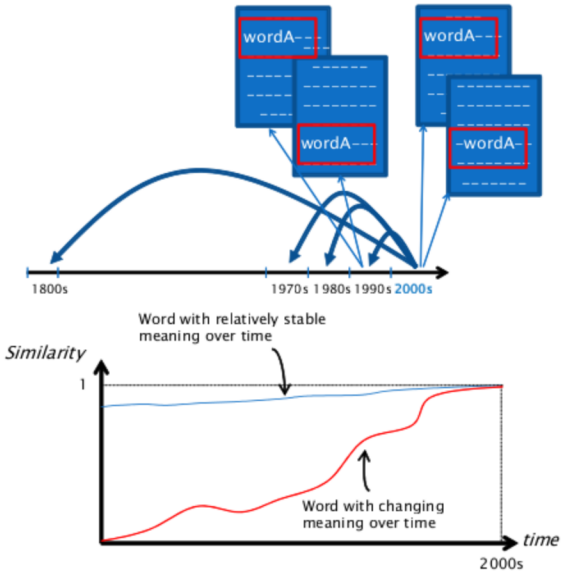


Figure 2 Concept of the past-present comparison of word contexts constructed from multiple texts across decades (top) and resulting similarity curves (bottom) for two hypothetical words.

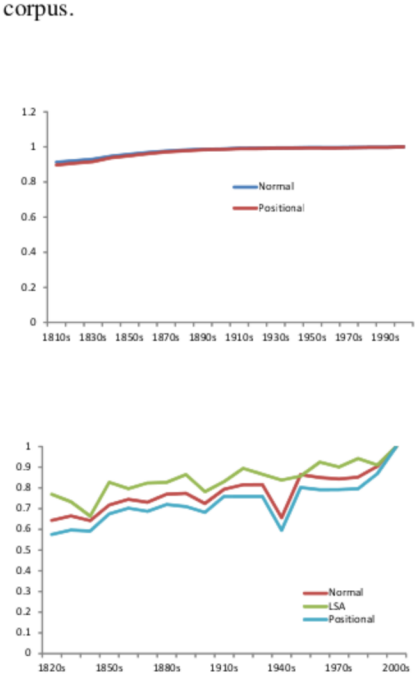
4.1.1 Case Study

As we mentioned before, words that have similar usage in different decades should have vector representations that are close by. On the top-left hand side of Figure 3 we can see the similarity curve computed for the example word "god" on Google Books 5-gram corpus. Clearly, the context of this word is almost same across decades. This conclusion is supported by the evidence in the form of the representative words for each decade (see the top-right hand side of Figure 3). The way to find these words will be described later. We suspect that the relatively high stability of this word's usage over time comes from many religious texts such as prayers or various editions of the Bible that kept their form more or less unchanged across decades. Additional evidence for this conclusion comes from nearly identical values of the similarity plots for both the normal and positional representations which indicate that the positions and order of words remained relatively stable over time.

For comparison, at the bottom of Figure 3 we show the results obtained on COHA dataset. All the three types of context representations are similar and are relatively stable. The representative words are however more varying compared to the

[10] Other measures such as Pearson correlation and Euclidean distance could be also used here.

ones derived from Google Books 5-gram dataset. This may be due to more selective and diversified way of constructing COHA corpus.



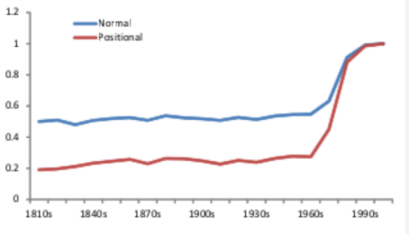
1820s	word	love	spirit
1830s	word	love	spirit
1840s	word	love	lord
1850s	word	love	lord
1860s	word	love	lord
1870s	word	love	lord
1880s	word	love	kingdom
1890s	love	kingdom	word
1900s	love	kingdom	word
1910s	kingdom	love	word
1920s	love	kingdom	word
1930s	love	word	kingdom
1940s	love	kingdom	word
1950s	love	word	kingdom
1960s	word	love	kingdom
1970s	love	word	kingdom
1980s	word	love	kingdom
1990s	word	love	kingdom
2000s	word	love	lord

1820s	love	grace	word
1830s	love	man	son
1840s	man	word	love
1850s	word	man	thank
1860s	thank	man	given
1870s	love	thank	bless
1880s	man	thank	word
1890s	love	thank	lord
1900s	name	mercy	thank
1910s	word	thank	knows
1920s	knows	help	thank
1930s	knows	thank	wish
1940s	lord	knows	thank
1950s	knows	thank	help
1960s	knows	thank	love
1970s	knows	thank	believe
1980s	knows	thank	name
1990s	knows	believe	thank
2000s	knows	thank	word

Figure 3 Past-present similarity plots and the representative words for the word "god" calculated on Google Books 5-gram (top) and on COHA (bottom) datasets.

On the other hand, looking at the example of the evolution of the word "gay" we can see the sudden increase in the similarity curve that occurred about three or four decades ago (see Figure 4). This is in accordance to the etymological explanation of the word "gay", which before used to mean: "…'excellent person, noble lady, gallant knight,' also 'something gay or bright; an ornament or badge'"... "Gay as a noun meaning 'a (usually male) homosexual' is attested from 1971…" [10].

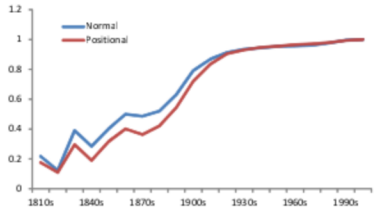
Indeed, looking at the associated words we can recognize that the current meaning of homosexuality came around 1970s or 1980s. Another interesting observation is the steady appearance of word "lussac" until 1960s. It refers to the French Chemist called Gay Lussac[11] who lived from 1778 to 1850.



1820s	lussac	world	young
1830s	lussac	world	young
1840s	lussac	world	young
1850s	world	lussac	young
1860s	lussac	world	young
1870s	lussac	grave	young
1880s	lussac	young	john
1890s	lussac	grave	young
1900s	lussac	grave	young
1910s	lussac	young	grave
1920s	lussac	life	grave
1930s	lussac	life	young
1940s	lussac	life	young
1950s	lussac	life	young
1960s	lussac	life	young
1970s	liberation	world	lussac
1980s	lesbian	lesbians	community
1990s	lesbian	lesbians	community
2000s	lesbian	lesbians	community

Figure 4 Past-present similarity plots (left) and the top representative words for the word "gay" (right) calculated on Google Books 5-gram dataset.

The word "propaganda" is another example. Its modern (usually negative) political sense dates from the World War I. Before it was used in association with "Congregatio de Propaganda Fide" (Congregation for Propagation of the Faith) set up by Pope Gregory XV in 1622 in order to spread the world of Christianity by missions around the world [4]. It also represented "any movement to propagate some practice or ideology" [10]. We show its similarity plot and the representative words in Figure 5.



1820s	lussac	world	young
1830s	lussac	world	young
1840s	lussac	world	young
1850s	world	lussac	young
1860s	lussac	world	young
1870s	lussac	grave	young
1880s	lussac	young	john
1890s	lussac	grave	young
1900s	lussac	grave	young
1910s	lussac	young	grave
1920s	lussac	life	grave
1930s	lussac	life	young
1940s	lussac	life	young
1950s	lussac	life	young
1960s	lussac	life	young
1970s	liberation	world	lussac
1980s	lesbian	lesbians	community
1990s	lesbian	lesbians	community
2000s	lesbian	lesbians	community

Figure 5 Past-present similarity plot (left) and the top words (right) for "propaganda" using Google Books 5-gram datasets.

4.2 Change Explanation

The representative words shown in Figures 3-5 are collected in the following way. First, for each decade we remove all context words that have frequency smaller than 1% of the most frequent word within the decade. By applying this filtering condition on the frequency we remove many noisy words that could be the result of OCR errors. Then the frequency of each remaining context word in decade d_i is compared to the frequency of the same word in decade d_{i-1}.

$$S(a,d_i) = f(a,d_i) \,/\, f(a,d_{i-1})$$

The top-scored words are then returned. These words are characterized by a large increase in the co-occurrence with the target word along the time flow from decade d_{i-1} to d_i.

In Figure 6 we show top context words for two examples of target words, "toilet" and "mouse". The first one meant a table used for face and hair arrangement for ladies around 19th century as well as the activity of getting oneself ready for a day [4,10]. Around the beginning of the last century, however, the word toilet started to assume the meaning of dressing room to later denote the bathroom. Looking at the terms returned for the word "mouse" we can understand that the meaning of animal was the likely meaning for most of the time within the last two centuries. The recent meaning of a computer mouse can be inferred through the words like "button", "click", "left" and "pointer". The words "cells" and "brain" are likely due to the common practice of using mice in laboratory experiments.

We propose treating the word lists as evidence for supporting the analysis and understanding of etymology-related statistics such as the past-present similarity plots introduced in the previous section. At the same time they constitute discovery tool for finding new word meanings and types of usage.

[11] http://en.wikipedia.org/wiki/Joseph_Louis_Gay-Lussac

1820s	her	table	lost	1820s	country	city	mouse
1830s	duties	table	his	1830s	white	country	trap
1840s	table	her	completed	1840s	field	country	white
1850s	table	make	her	1850s	country	white	field
1860s	table	made	make	1860s	country	trap	field
1870s	table	made	make	1870s	country	field	trap
1880s	table	made	make	1880s	country	trap	field
1890s	table	made	make	1890s	trap	country	field
1900s	table	made	articles	1900s	trap	country	white
1910s	rooms	articles	table	1910s	trap	country	mouse
1920s	articles	table	rooms	1920s	trap	field	white
1930s	articles	table	facilities	1930s	trap	field	church
1940s	paper	facilities	articles	1940s	house	mouse	trap
1950s	training	paper	preparations	1950s	house	game	trap
1960s	training	preparations	paper	1960s	cells	house	game
1970s	training	paper	seat	1970s	cells	brain	cell
1980s	paper	training	seat	1980s	cells	button	mammary
1990s	paper	seat	bowl	1990s	button	left	pointer
2000s	paper	seat	bowl	2000s	button	pointer	click

Figure 6 Top growing context words for "toilet" (left) and "mouse" (right) obtained from Google 5-gram dataset.

4.3 Inter-decade Word Similarity

In Section 4.1 we introduced a particular way of looking at word meaning evolution in which present context of the word is contrasted with its past contexts. Here we calculate the similarities of word context across all possible decades to generate more thorough visual representation of shifts in word meaning. Figure 7 and 8 show generated views for the example words "Halloween" and "atomic" in the form of colour matrix complemented with representative words. Red colour indicates high contextual similarity for a given pair of decades while green represents low similarity. The decades are indicated here as numbers ranging from 1 to 20. "Halloween" as a festival is known to have been popularized within United States at the beginning of 20th century after having been brought by Scottish immigrants [21]. From that time on, the meaning of party or special night was becoming commonly used. This is evidenced by the square pattern that is continuously indicated in red and that starts from 11th decade (1910s). The pattern is associated with the representative words "party" and "night" as shown inside the table on the right hand side of Figure 7.

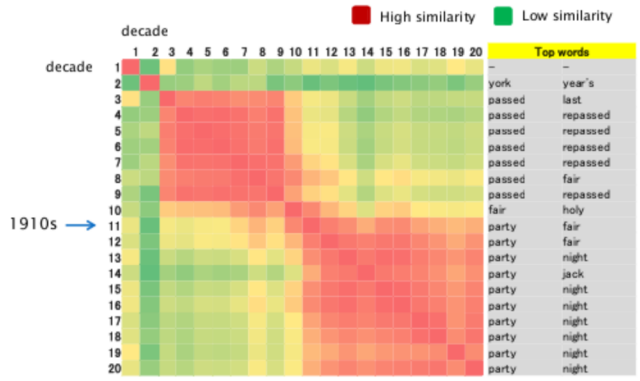


Figure 7 Inter-decade matrix and top representative words for target word "halloween" calculated on Google Books 5-gram dataset using normal vector representation.

Looking at Figure 8 we can see similar situation. There is contiguous area of high inter-decade similarity for the word "atomic" that covers period from 1940s to the last decade. As it is easy to guess "atomic" was started to be commonly used in the context of bomb or energy following the tragedy of Hiroshima and Nagasaki bombing at the end of the World War II. The newly acquired meaning triggered by the sudden event immediately dominated other meanings. This can be observed by comparing the shapes of squares in Figure 7 and 8.

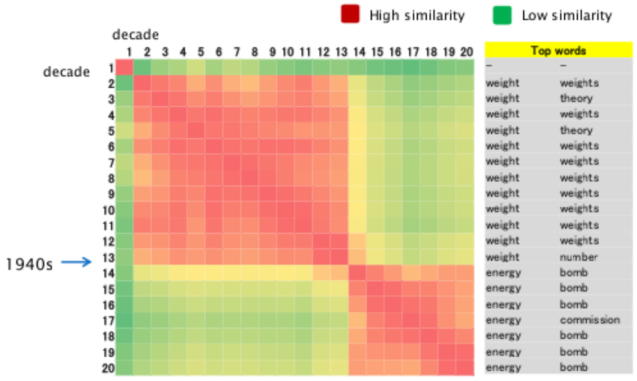


Figure 8 Inter-decade matrix and the top representative words for target word "atomic" calculated on Google Books 5-gram dataset using normal vector representation.

4.4 Evolution of Contrastive Word Pairs

Sometimes we wish to know how the similarity between two words changed over time. For example, it may be easier to understand the change in the meaning of one word by comparing it over time with other words that are supposed to have similar meanings. We thus propose comparing representations of two words in each decade. Figure 9 portrays the concept of synchronous context comparison between two words.

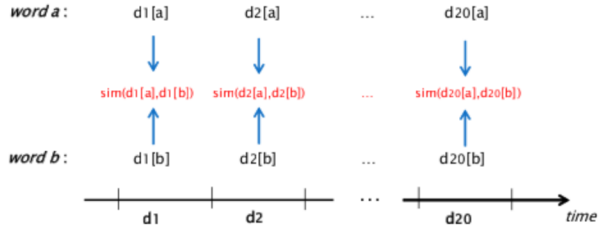


Figure 9 Concept of vector comparison across time for a word pair input.

Such an across-time word-word similarity calculation could be then another mechanism for better quantifying the actual meaning change of target words. For example, it is known that the word "nice" and "pleasant" become similar in meaning over time. "Nice" is actually an extraordinary example of adjective as it had variety of meanings in distant past. The previous sense of "nice" was close to "foolish", "silly", "ignorant" or "cowardly", to become later to mean "strange", "rare" and "reserved" [4]. The sense of "agreeable", "delightful" was recorded from 1769 and the one of "kind", "thoughtful" from 1830 [10]. On the other hand the word "pleasant" maintained rather similar meaning over time [10]. On Figure 10 the plot indicates the convergence in meanings that took mostly place from the middle of 19th century. Each point in the plot indicates the similarity of the context vector of "nice" and "pleasant" at given time.

Similarly, it is known that the word "guy" become close in meaning to the one of "fellow" around the end of 19th century following the publication of popular book titled, "Guy Fawkes" by W. H. Ainsworth in 1840s about the life of the historical character Guy Fawkes [4,10]. We can verify this fact when looking at Figure 11.

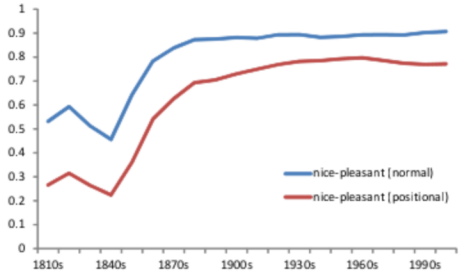


Figure 10 Similarity plots for word pair: "nice" and "pleasant" calculated on Google Books 5-gram dataset.

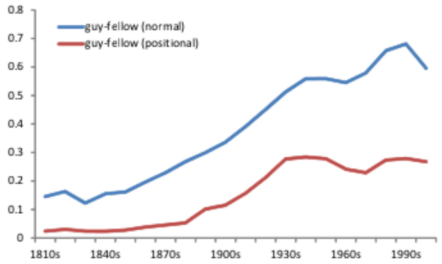


Figure 11 Similarity plots for word pair: "guy" and "fellow" calculated on Google Books 5-gram dataset.

The across-time word to word meaning comparison should be even more informative when coupled with the information on the semantic change of the compared words and their top representative context words as described in the previous sections. For example, we show below the plot comparing similarities of "mouse" and "rat" over time. The drop in similarity starting from around 1970s could be attributed to the new usage of "mouse" to refer to a computer device. The noticeable divergence of the normal and positional curves at that time further confirms the shift in the context. This validates the findings from the Figure 6 in which top representative words point to the acquisition of another meaning of "mouse" around 1960s-1970s.

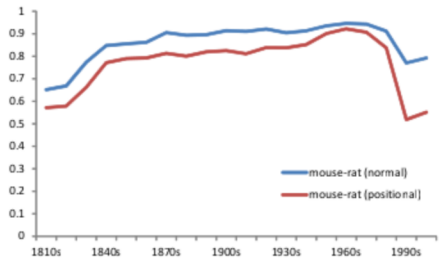


Figure 12 Similarity plots for word pair: "mouse and "rat" calculated on Google Books 5-gram dataset.

The other examples we show are for word pairs "propaganda" vs. "information" in Figure 13 and "toilet" vs. "washroom" depicted in Figure 14. They confirm our previous findings described in Sections 4.1 and 4.2.

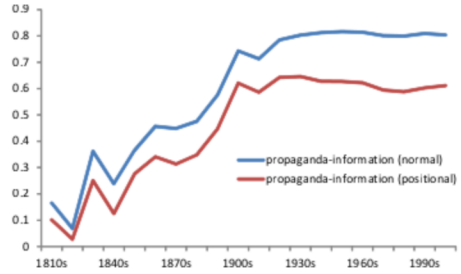


Figure 13 Similarity plots for word pair: "propaganda" and "information" calculated on Google Books 5-gram dataset.

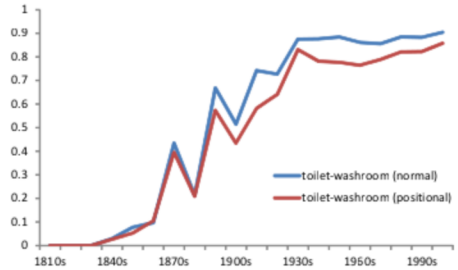


Figure 14 Similarity plots for word pair: "toilet" and "washroom" calculated on Google Books 5-gram dataset.

4.5 Sentiment Analysis over Time

Each word has certain sentiment value that can vary over time. We measure such value by the evaluation of the sentiment of its context words in different decades. We use here positivity and negativity scores of words as provided in SentiWordnet[12], which is lexical resource for opinion mining. Since the sense information of the context words is not given we use the weighted average positivity and negativity scores over different senses of each word. The weights are linearly correlated with the ranks of word senses as indicated in SentiWordnet.

In essence, we rely here on majority information coming from multiple words to provide a reliable signal at each time unit. Note that we assume here stable sentiment values of context words over time. By doing this the calculation becomes simpler. This assumption is fine considering that among relatively large number of context words the sentiment changes of some individual words will have little effect on the overall results. Future approaches could experiment with sentiment vocabulary containing only words that had relatively stable sentiment values over the past decades.

Finally, we plot the average positive and negative scores of the context words in each decade as well as the total sentiment which is the difference of the negativity score and the positivity score at every decade.

Figure 15 shows the sentiment plot for the word "aggressive" which is known to become less negative in recent times than in the past. As stated in [12]: "…In more recent times, the steady amelioration of ambitious and aggressive reveals a change in attitude towards those who seek advancement or 'success' in a highly competitive fashion…." Indeed, the amelioration process of this word becomes evident when looking at decreasing value of the negativity curve.

[12] http://sentiwordnet.isti.cnr.it

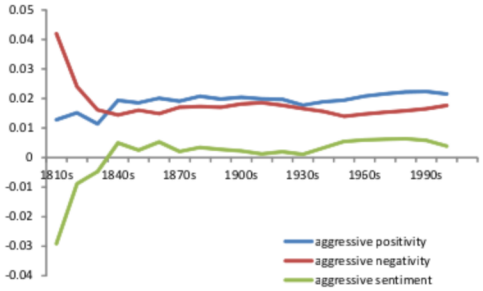


Figure 15 Across-time sentiment analysis of word "aggressive" calculated on Google Books 5-gram dataset.

To obtain more insight into the change of the meaning of "aggressive" we show in Figure 16 the top representative words taken from Google Books 5-gram (left-hand side) and from COHA (right-hand side). We can notice that the usage of "aggressive" has become more common in positive situations as indicated by the appearance of words "policy", "growth", "fund" or "campaign" and disappearance of words like "tyranny", "war", "warlike" or "hostilities". We then provide more evidence by also showing in Figure 17 the similarity plot between "active" and "aggressive" as derived from Google Books 5-gram corpus. The similarity curve has the largest increase more or less around the same time when the decrease in negativity occurs as it could be seen in Figure 15.

1820s	tyranny	assisting	france	1820s	ways	way	warlike
1830s	hostilities	tyranny	think	1830s	ways	way	warlike
1840s	hostilities	forces	seem	1840s	foreign	ways	way
1850s	spirit	acts	policy	1850s	power	warlike	force
1860s	policy	spirit	movement	1860s	policy	warlike	ways
1870s	policy	movement	make	1870s	power	designs	action
1880s	policy	attitude	movement	1880s	policy	force	assume
1890s	policy	attitude	spirit	1890s	character	policy	campaign
1900s	policy	attitude	part	1900s	policy	campaign	positive
1910s	policy	action	attitude	1910s	policy	little	policies
1920s	policy	action	attitude	1920s	action	threatened	campaign
1930s	policy	action	attitude	1930s	wage	leadership	result
1940s	policy	action	attitude	1940s	make	right	policy
1950s	policy	behavior	action	1950s	policy	start	used
1960s	behavior	policy	impulses	1960s	behavior	tactics	members
1970s	behavior	policy	sexual	1970s	ambitions	israel	stance
1980s	behavior	policy	sexual	1980s	deliberate	ways	start
1990s	behavior	policy	passive	1990s	growth	fund	little
2000s	behavior	policy	passive	2000s	fight	waged	approach

Figure 16 Top representative words for "aggressive" selected from Google Books 5-gram (left) and COHA (right) datasets.

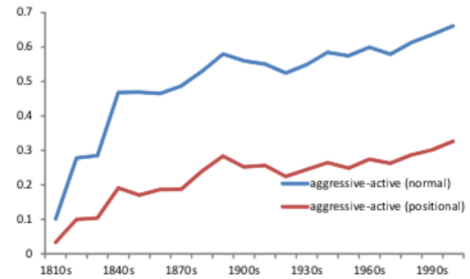


Figure 17 Similarity plots between "active" and "aggressive" based on Google Books 5-gram dataset.

As the next example, Figure 18 shows the sentiment plot for the word "propaganda". We can observe the pejoration process starting from the beginning of the last century which correlates with its meaning change as evidenced by Figure 5.

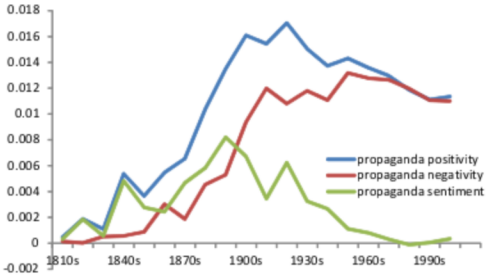


Figure 18 Across-time sentiment analysis of word "propaganda" calculated on Google Books 5-gram dataset.

As the last example we show in Figure 19 the results for the word "fatal" that meant "determined by fate" in the past; however, it has been recently used to mean "causing death" [10]. We observe indeed the gradual increase in the negative sentiment of this word. When we contrast the plot in Figure 19 with the top representative terms (see Figure 20) we notice the appearance of words that reveal clue expressions for this change such as "fatal disease", "fatal error", "fatal shooting" or "fatal mistake" over the last century.

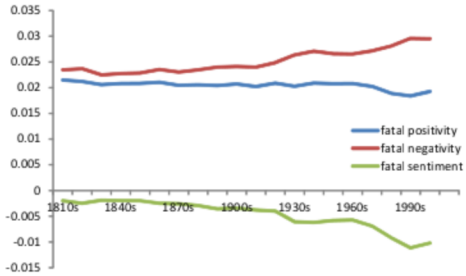


Figure 19 Across-time sentiment analysis of word "fatal" calculated on Google Books 5-gram dataset.

1820s	proved	consequences	blow	1820s	proved	blow	prove
1830s	proved	consequences	prove	1830s	proved	blow	consequences
1840s	proved	consequences	blow	1840s	proved	night	blow
1850s	proved	consequences	prove	1850s	blow	would	day
1860s	proved	blow	prove	1860s	proved	prove	might
1870s	proved	prove	blow	1870s	mistake	blow	proved
1880s	proved	prove	blow	1880s	blow	proved	mistake
1890s	proved	prove	blow	1890s	blow	proved	prove
1900s	proved	prove	blow	1900s	mistake	proved	prove
1910s	cases	prove	proved	1910s	proved	day	blow
1920s	prove	proved	mistake	1920s	mistake	made	blow
1930s	prove	proved	mistake	1930s	mistake	proved	made
1940s	prove	cases	mistake	1940s	blow	error	proved
1950s	prove	cases	proved	1950s	mistake	shooting	disease
1960s	prove	cases	blow	1960s	prove	shooting	likely
1970s	prove	disease	blow	1970s	mistake	made	disease
1980s	disease	prove	blow	1980s	mistake	shooting	heart
1990s	disease	blow	mistake	1990s	heart	attack	shooting
2000s	blow	mistake	proved	2000s	mistake	potentially	accident

Figure 20 Representative words for "fatal" calculated on Google Books 5-gram (left) and COHA (right) dataset.

5. DISCUSSION

In this section we provide general discussion of important features characterizing our approach, we outline possible paths of further extensions and also list shortcomings of this work.

5.1 General Aspects and Future Directions

5.1.1 Usefulness

We believe that the results of word evolutionary analysis could aid in verifying hypotheses posted by linguists, anthropologists or other researchers. Although etymological dictionaries already exist, they often only refer to selected words as well as they may inform about only the first recorded meaning of words rather than providing the whole and detailed spectrum of word evolution. Thus

automatic analysis should offer advantage here, too, as returning more complete and comprehensive information. For instance, it should be also possible to automatically recommend words with particular types of temporal changes so as they could be later manually analysed.

5.1.2 Multi-view Framework

Since available data for certain time periods can be sparse and also precisely tracking something as vague as the evolution of word meaning is inherently difficult, we need different kinds of evidence to be able to reach any judgments. Thus we use several pieces of information together in a framework that helps with "assembling the puzzles". Corroborating results across different approaches, albeit simple ones, is necessary for the kind of data we operate on. Our framework advocates exploration of word semantic change at the lexical level, at the contrastive-pair level, and at the sentiment orientation level. Also, the three kinds of word representation are used to capture as much information about word context as possible (frequency, position and relation to other words). Although the plot shapes for these three representations can be similar, their relative values may differ depending on target words and time periods. For example, for "propaganda" (Fig. 5) the normal and positional representations yield closely spaced plots (less than 25% value difference between both plots), while for the word "gay" (Fig. 4) we observe significant differences (over 100% value difference). This suggests the contextual words used in the former case were generally appearing at same positions with respect to the target word, while for the latter the word positions changed much providing more confirmation towards the conclusion of the strong meaning change of "gay".

5.1.3 Cross-corpora Comparison

To better assist users in inferring conclusions on word evolution we propose synchronizing results from two corpora that have different characteristics. We note that such comparative analysis could be also interesting when two or more languages are compared. Actually, the so-called transition problem [27] is one of the key five foci in the studies of historical linguistics and it refers to the transfer of a linguistic variable from one language to another through the diffusion or propagation processes. As Google Books n-gram datasets are available for several different languages (French, German, English, US English, etc.) it is then possible to compare the popularity and meaning of words across two dimensions: time and language. For example, it should be feasible to compare the usage of the same word in British English over time with its usage in American English, or with even more distant language like French. This would be useful for finding cases where words have been borrowed from another language. In fact the inter-language word borrowing is one of key drivers of the language evolution [14,24] (consider numerous loan words that migrated from French to English such as "entrepreneur" or "rendezvous"). We emphasize that our methodology is language-independent and can be applied to different languages with only minor tweaks.

5.1.4 Concept Level Analysis.

The current approach could be also extended to the concept-level investigation. The problem of studying the concept evolution is more difficult than the analysis of single word's etymology. However, it would more accurately reflect the rise and fall of the popularity of entire concepts over time and the fluctuations of their actual meanings. Thus the study of the temporality and dynamics of language should not be limited to independent words only. For example, to accurately portray the history and the evolution of the concept "car" one could track not only the changes in the usage of the term "car" itself but also ones in its synonyms (e.g., "auto", "automobile", "vehicle") as well as one could incorporate semantic changes of other related or contrastive concepts (e.g., "cart", "bicycle", "train", "highway", "wheel", "motor"). In this process any known semantic interrelations between words should be considered (e.g., meronymy, anotonymy, etc.). Similarly, to properly reflect the popularity of the concept, "programming languages", and its meaning over time we could combine temporal characteristics of words referring to each individual programming language.

5.2 Limitations

It is quite good that the proposed framework is giving interesting results despite being relatively simple. For a large scale datasets that we use simple approaches seem to be currently the best option. Below we list the limitations of the current approach.

5.2.1 Different Word Meanings

Although we are able to track changes in dominant meaning of a term, it is still not possible to quantify the strength of the entire spectrum of meanings the word associates with at any given decade. Thus the framework cannot inform exactly to what extent each particular meaning of a word existed at a certain time point.

Subtle semantic changes. One of the issues that we noticed when closely analysing the results is related to the change from one meaning to a similar one (e.g., "corn" used to mean grain and wheat to be later used for referring to maize [10]). This kind of subtle changes may be difficult to detect due to relatively coarse representation of word meaning.

5.2.2 Spelling Variations.

We note that we did not care about spelling variations in this work. This might become an important issue to tackle when doing analysis spanning longer periods of time. Books from distant past can contain diverse surface forms of the same words, especially, named entities. The different spelling variations of words should be then tracked and their context should be merged.

5.2.3 Semantic Change of Context Words

Another issue that we discuss here is related to the semantic change of the context words. We have already mentioned it briefly in the section on the sentiment analysis. With the current method we essentially assume stationary character of the context semantics. A future extension of the current approach should however consider the semantic change of its context words instead of blindly treating them as static across time. This would necessarily result in a recursive problem of the meaning evolution that is not trivial, especially given the large size of data.

5.2.4 Rare Words

An obvious problem that should be mentioned relates to rare words for which there may not be enough context data available in certain time, especially, when the word was not commonly used. Combining data from different corpora could help to some extent.

6. CONCLUSIONS

"Every word has its own history." This motto has driven much of our research. Each word is different, and it has been challenging for the historical linguistics community to formulate wide-encompassing generalizations about lexical change [1,24]. We believe that the computational methods based on large diachronic corpora have significant potential to automatically discover interesting phenomenon for further manual studies.

Our contribution here is a visual analytics framework for discovering and visualizing lexical change at three different levels—individual words, word pairs, and sentiment orientation. We demonstrate several algorithms for finding and understanding the meaning change over time based on the Google Books corpus, which is the largest available historical dataset and on more

balanced COHA corpus. We then corroborate the findings with the state-of-the-art etymological knowledge showing that the automatic analysis could assist in manual explorations.

Besides the already mentioned future work, we plan to carry more extensive quantitative analysis. The first step is to prepare a labelled test set so that we can make rigorous comparisons and infer quantitative conclusions. Based on the current work, we also seek to build an online service with a rich web interface allowing insightful queries over massive historical textual data. The planned service should foster linguistics studies and allow general public to inquire about evolution of interesting concepts and to appreciate the richness of their language. The final step is to study deeper the actuation problem [27] behind semantic changes to detect not only the meaning transitions themselves but also their reasons and any underlying forces that triggered them.

7. ACKNOWLEDGMENTS

This research was supported in part by Ministry of Education, Culture, Sports, Science, and Technology (MEXT) Grant-in-Aid for Young Scientists B (#22700096) and by the Japan Science and Technology Agency (JST) research promotion program Sakigake: "Analyzing Collective Memory and Developing Methods for Knowledge Extraction from Historical Documents".

8. REFERENCES

[1] J. Aitchison, *Language Change, Progress or Decay?* Cambridge University Press, 2001.

[2] L. Campbell, *Historical Linguistics*, 2nd edition, MIT Press, 2004.

[3] B. Castle, *Why Do We Say It? The Stories Behind the Words, Expressions and Cliches We Use*. (Reissue edition) 1986.

[4] J. Cresswell, *Dictionary of Word Origins*, Oxford Press, 2010.

[5] M. Davies, "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English," in *Literary and Linguistic Computing*, 25(4):447–464, 2010.

[6] S. Deerwester et al., "Improving Information Retrieval with Latent Semantic Indexing," in *Proceedings of the 51st Annual Meeting of the American Society for Information Science*, 25, pp. 36–40, 1988.

[7] J. Firth, "A synopsis of linguistic theory 1930-1955," in *Studies in Linguistic Analysis*, Oxford Philological Society, pp. 1–32, 1957.

[8] A. Gerow, and K. Ahmad, "Diachronic Variation in Grammatical Relations," in *Proceedings of COLING 2012*, pp. 381–390, 2012.

[9] Y. Goldberg and J. Orwant, "A Dataset of Syntactic-Ngrams over Time from a Very Large Corpus of English Books," in *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pp. 241–247, 2013.

[10] D. Harper, *Online Etymology Dictionary*. http://www.etymonline.com/ (compiled from over 70 sources including major etymological dictionaries; data retrieved on 14th March, 2014)

[11] Z. Harris, "Distributional Structure," in *Word*, 10(23):146–162, 1954.

[12] G. Hughes, *Words in Time: A Social History of the English Vocabulary*. Basil Blackwell, 1988.

[13] A. Jatowt, and K. Tanaka, "Large Scale Analysis of Changes in English Vocabulary over Recent Time" in *Proceedings of CIKM 2012*, pp. 2523-2526, 2012.

[14] W. Labov, *Principles of Linguistic Change (Social Factors)*, Wiley-Blackwell, 2010.

[15] A. Lehrer, *Semantic Felds and Lexical Structure*. North-Holland, American Elsevier, Amsterdam, NY, 1974.

[16] A. Liberman, *Word Origins And How We Know Them: Etymology for Everyone*. Oxford University Press, 2009.

[17] C. Mair, et al., "Short Term Diachronic Shifts in Part-of Speech Frequencies: a Comparison of the Tagged Lob and F-lob Corpora," in *International Journal of Corpus Linguistics*, 7(2):245–264, 2003.

[18] J.-B. Michel, et al., "Quantitative Analysis of Culture Using Millions of Digitized Books" in *Science*, 331(6014), pp. 176-182, 2011.

[19] R. Mihalcea, and V. Nastase, "Word Epoch Disambiguation: Finding How Words Change Over Time" in *Proceedings of ACL (2) 2012*, pp. 259-263, 2012.

[20] S. Mohammad, "From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 105–114, 2011.

[21] R. Nicholas, *Halloween: From Pagan Ritual to Party Night*. New York: Oxford Univ. Press, 2002.

[22] D. Odijk et al., "Exploring Word Meaning through Time," in *Proceedings of the TAIA'12 Workshop in conjunction with SIGIR'12*. Portland, USA, 2012.

[23] J. Platt, K. Toutanova, and W.T. Yih, "Translingual Document Representations from Discriminative Projections," in *Proceedings of EMNLP 2010*, pp. 251-261, 2010.

[24] H. Schendl, *Historical Linguistics*. Oxford University Press, 2008.

[25] N. Tahmasebi et al., "NEER: An Unsupervised Method for Named Entity Evolution Recognition," in *Proceedings of the Coling'12*, ACL, pp. 2553-2568, 2012.

[26] P. Turney, and M. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," in *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[27] W. Uriel et al., "Empirical Foundations for a Theory of Language Change," in *Directions for Historical Linguistics*. Austin, London, University of Texas Press, pp. 95-188, 1968.

[28] G. K. Zipf, *Human Behavior and the Principle of Least Effort*. Cambridge, (Mass.): Addison-Wesley, 1949.