Large Scale Analysis of Changes in English Vocabulary over Recent Time

Adam Jatowt[1,2] and Katsumi Tanaka[1]
[1]Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan
[2]Japan Science and Technology Agency
4-1-8 Honcho, Kawaguchi-shi, Saitama
332-0012 Tokyo, Japan
{adam, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT
Recently many historical texts have become digitized and made accessible for search and browsing. As human language is subject to constant evolution, these texts pose varying challenges to current users. In this paper we report the results of large-scale studies on the usage of words and the evolution of English language vocabulary over the last two centuries to help with understanding its impact on readability and retrieval of historical documents. We perform analysis of several lexical factors which may influence accessibility and readability of historical texts based on two large scale lexical corpora: the *Corpus of Historical American English* and *Google Books 1-gram*.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords
Language evolution, historical texts, readability, information retrieval

1. INTRODUCTION
In the course of recent years we have witnessed massive digitalization of historical texts carried by libraries, museums and numerous other institutions. To comply with preservation and accessibility objectives, many old books, news articles, letters and other documents have been scanned, subject to optical character recognition and made available to public. Project Gutenberg (www.gutenberg.org), Google Books (http://books.google.com) and Internet Archive Text Collection (www.archive.org) are examples of such initiatives. For the first time large amounts of historical texts have been made available for users.

Common sense tells that old documents are difficult to read due to different vocabulary, obsolete patterns of word usage, different grammatical structures and other factors resulting from time passage. Scientists have put hypotheses of colloquialization and democratization of language in the recent centuries and decades proving that on average texts became progressively easier [8,10]. For example, it has been found that written sentences became shorter over time [10]. This situation is contributed to the fact that in the past, generally, mainly well-educated people could write and publish while nowadays these restrictions have lesser importance. Jatowt and Tanaka [6] have confirmed negative correlation between documents' age and their readability, as experienced by current readers, using standard readability metrics based on sentence- and word-length. However, document readability is based on many other constructs. To better understand readability issues related to historical texts, more analysis is needed, especially the one based on comparing vocabulary usage across time. Apart from the readability problem, the language change has also impact on how effectively users can retrieve old documents or contained information by using free text queries. Intuitively, since the current readers and the past authors lived in distant times, there is certain difference in their operational vocabularies, which may then impact the reading and retrieval processes that the current users undertake.

Although the language evolution is a known fact, few previous works have tried to measure it from macro-scope viewpoint using large-scale datasets. Moreover, no empirical studies focused explicitly on the ease of reading and retrieval of historical documents in connection to language evolution, even though, they are necessary to understand barriers and cognitive loads imposed on current users. In this paper we attempt at filling these gaps using large diachronic datasets.

We investigate the scope of vocabulary change in English across the last two centuries and its related characteristics in order to better understand the impact of time on document readability and accessibility. The questions that we concentrate on are "how much the active vocabulary changed over the recent time?" and "how this change may affect current users who wish to read and retrieve old documents?" We focus on the last two centuries as they embrace the time period during which the majority of historical texts that are stored by libraries and archival institutions were created.

The results of our analysis could be also helpful for improving OCR recognition and can enhance methods for dating historical documents [3,7]. The latter has applications in generating or verifying metadata of historical texts. The usual approaches rely on employing features such as temporal language models, diachronic frequencies of words or occurrence of named entities. We think that the findings and some features discussed in this paper such as unique word probability per decade, changes in word length, POS tag distributions over time or temporal entropy/kurtosis of words could become also useful for these tasks.

2. DATASETS
In order to study the language evolution we need datasets large enough to support drawing valid conclusions. We use two lexical

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

corpora, *Corpus of Historical American English* (*COHA*)[1] and *Google Books 1-gram*[2]. COHA contains over 400 million words collected from about 107K documents published from 1810s to 2000s. The documents were carefully selected by keeping the same ratio of different genres throughout different decades according to the Library of Congress categorization system. In terms of the OCR error rate the corpus is 99.85% accurate, which translates to one error for about every 500-1000 words.

COHA is divided into 20 decades, and the frequency of each word is reported for every decade. On average there are 20.2 million words per decade. We have removed data for the first decade as in most of the cases it generates outlier values inconsistent with the rest of the data. This is attributed to relatively low total word count for this decade when compared to other decades (about 20 times smaller size than the average). COHA contains also part-of-speech (POS) tags indicating grammatical roles of terms in their original texts and the calculated total frequency of each word-POS tag pair in every decade. We will use this feature in Section 3.3.

The second dataset, the Google 1-gram, is much larger as it has been compiled from the Google Books project which claims to contain about 5% of books ever published. The data on term frequency is available for each year for the last two centuries. For the comparison purpose with the COHA dataset we converted the yearly granularity to the decade granularity within the period from 1820s to 2000s. On average, the dataset contained 17.9 billion words per decade. Other preprocessing steps involved converting words to small cases and removing digits and other non-words. Google n-gram datasets were used for *culturonomics* [9] which is a study of the changes in word usage and cultural trends over time. However, unlike our work that study was rather focused on individual words or their sets.

We applied a threshold on the frequency of words in each decade equal to 50 words for COHA and 300 for Google 1-gram to remove tokens generated as a result of OCR errors or ones specific only to a particular document or author. These thresholds were applied in all the studies reported throughout this paper.

Note that both datasets are substantially different. Google 1-gram has been generated based on all the digitized books within the Google Books initiative, while COHA contains carefully selected and balanced prose texts with a relatively stable rate of diverse genres across different decades. Thus their simultaneous usage makes sense, however due to space constraints we sometimes report the results obtained from only one dataset.

3. ANALYSIS

3.1 Number of Unique Words across Time

We first focus on changes in vocabulary distributions over recent decades in order to capture differences in word usage and the emergence of new words. Vocabulary burden in text comprehension was already observed by linguists long time ago [4]. High density of unknown words in text slows down reading process, increases cognitive burden and impairs information recall. Our objective is to quantify the divergence between words used by authors and readers who lived in different times. Figures 1 and 2 show the distributions of unique terms over time for COHA and Google 1-gram datasets, respectively. We can observe the increase in the number of unique words over time in both the datasets (37% total increase for COHA and 91% for Google 1-gram). We note that of course every word has its sense(s) that might be also changing over time. However, the analysis of the change of words' semantics is outside of the scope of this paper.

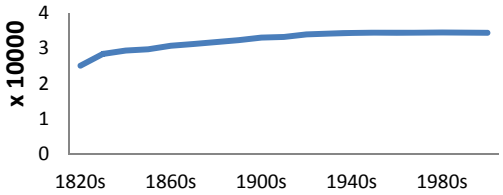

Figure 1 Number of unique words in COHA.

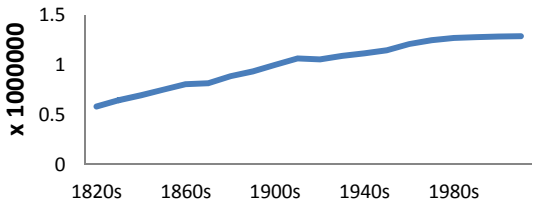

Figure 2 Number of unique words in Google 1-gram.

To consider the effect of different sizes of data in different decades we normalized the number of unique words by the total number of words in each decade. The resulting plot is shown in Figure 3. It can be interpreted as a kind of type-to-token ratio or lexical variety for every decade. It decreases -67% for COHA and -95% for Google 1-gram over the whole time span. Looking at the above results only one could say that the vocabulary appears to be more varied in texts published in distant past than in recent times, although to make sure we should compare text samples from different times. Note that this measure considers only the raw number of unique terms and does not take into account whether the words are known to current readers. In Section 3.3 we estimate the difference between vocabulary distribution of current English and those of older decades.

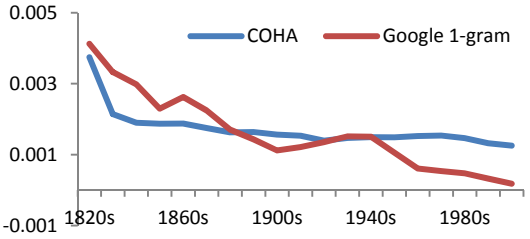

Figure 3 Average type-to-token rates in each decade in COHA and Google 1-gram.

We next calculated the mean frequency of words across all the decades. Such *average across-time frequency* indicates "persistence" of words over time. Persistent words may be considered as a sort of "bridge" over vocabulary gap between readers and writers living in different times since they are common to both groups. Table 1 lists 30 top words extracted from COHA dataset according to their average across-time frequency. The listed words are the ones performing basic grammatical functions such as articles, conjunctions and pronouns, hence, terms of little discriminative power that are usually regarded as stop words in IR.

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

[2] http://books.google.com/ngrams/datasets

Table 1 Top frequent words across decades in COHA (ordered from top to bottom and left to right).

the	that	is	you	at	have
of	i	his	on	but	do
to	he	for	had	her	they
a	it	with	be	she	this
in	was	as	not	by	from

3.2 Word Stability across Time

To gain more insight into the actual changes in the number of unique words over time, we grouped words into 4 bins according to their frequency within each decade. Since decades usually have varying number of total terms, thus a fixed threshold levels would not work. We instead applied frequency-based relative threshold which allows adapting the grouping to the sizes of word distributions in each decade. Let M be the frequency of the most frequent word in a decade, $M = max\ \{ln(F_i)\}$, where F_i denotes the frequency of arbitrary word i. Then a frequency bin is defined as:

$$M * \alpha \le \ln(F_i) < M * \beta \qquad (1)$$

It contains words whose frequencies are bounded by α% and β% of the frequency of the top-frequent word on a log scale ($\alpha \le 1$, $\beta \le 1$). For example, the bin 0%-25% contains 25% of the least frequent words in a given decade.

In Figure 4 we show changes in the number of unique terms within the 4 bins: 0%-25%, 25%-50%, 50%-75% and 75%-100% over different decades for COHA dataset. We can notice that the previously reported increase in the number of unique words appears to be mostly due to the growing number of words in the middle low part of the frequency spectrum (25%-50%).

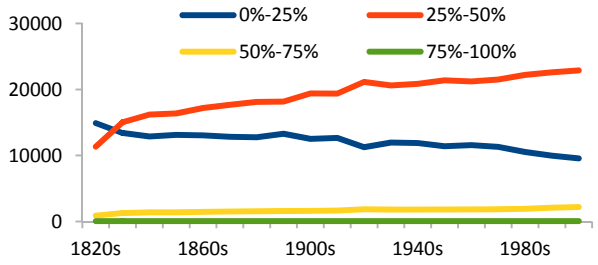

Figure 4 Number of unique terms over time in COHA grouped by their within-decade frequency.

We can interpret these results as the rich-get-richer phenomenon considering the previous finding that the number of unique terms increases along with time. In general, the rich-get-richer effect, known as *Matthew effect*, describes the process in which the choices made in the past are more likely to be selected also in the future. In terms of language evolution it would mean that words frequent in one decade tend to remain frequent in the subsequent decades. We suspect the newly created or absorbed words tend to fall mainly into the end or into the middle of the tail of the Zipfian distribution over words' popularity before they gradually become more popular. The number of unique words in the top part of the frequency spectrum (i.e., the most frequent words) changes too, but it changes less, in absolute terms, than the numbers of unique words in the lower parts of the spectrum (i.e., less frequent words).

To look from different angle on this hypothesis we have also analyzed entropy of term frequency distribution over different decades in COHA. The higher the entropy, the more stable a word is across time, meaning that its frequency changes less over decades. We found a weak positive correlation of *0.33* (Spearman's Rank Correlation Coefficient) between the temporal entropy of words and their average across-time frequency. Table 2 shows the top words ranked by their across-time entropy in COHA dataset. Note that the words in Table 2 were selected based only on their temporal entropy scores without considering their average across-time frequency.

Similarly, we also investigated kurtosis of the frequency plots which is often used as a measure of "peakness" of probability distribution. We found a *-0.24* correlation between the kurtosis and the average across-time frequency for words in COHA. From the language evolution's viewpoint, both of the above correlations imply that infrequent words are subject to much stronger variation over time than the frequent words confirming the rich-get-richer phenomenon in language. From the information retrieval viewpoint, the fact that the most popular words tend to remain popular, means that current users will be able to easily come up mostly with common words when constructing their queries to retrieve past documents. However, these are not the most efficient terms in IR as such terms usually poorly discriminate documents.

Table 2 Top words according to temporal entropy in COHA (ordered from top to bottom and left to right).

in	longer	when	directly
to	where	difference	the
other	for	more	it
an	that	if	they

3.3 Difference of Word Distributions

We measure here change in word distributions over time as a means of quantifying the overall differences between the English used at different decades. For this we utilize Jensen-Shannon (JS) divergence, which is a symmetric measure of difference between two distributions bounded by 0 and 1. Figure 5 shows the values of JS divergence between the last decade (fixed) and every previous decade for both the COHA and Google 1-gram datasets. We can observe that the divergence appears to grow linearly along with the time distance between the compared decades. The plots of the divergences have similar shapes for both the datasets, although the divergence on the Google 1-gram dataset has higher values, probably due to its much larger size. We note the steady, progressive nature of the change in word distributions across decades. The JS divergence between the vocabulary in 2000s and the one in 1820s is about 18 times higher for COHA and 22 times higher for Google 1-gram dataset than the corresponding JS divergences between the vocabulary in 2000s and the one in 1990s.

As mentioned above, COHA dataset contains pre-computed part of speech tags for each token together with its POS-dependent frequency in every decade. We use this data for comparing the total POS tags' distributions over time as a rough approximation of the grammar change over decades. Naturally, relying only on the grammatical functions of words, as indicated by POS tags, causes information miss such as the information on the word's typical order and function in text. Nevertheless, the grammatical functions of words indicate their key roles in text and their divergence can be used as one type of simple approximation of grammatical difference across time. POS-based JS divergence between the last decade and all the other decades in COHA is shown in Figure 6. It appears to have similar shape to the one for

the vocabulary-based JS divergence but is characterized by an order of magnitude smaller values.

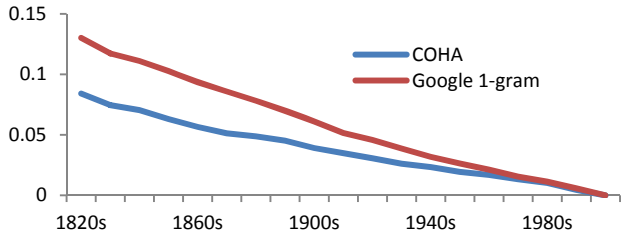

Figure 5 Vocabulary-based JS divergence between the last decade and other decades in COHA and Google 1-gram.

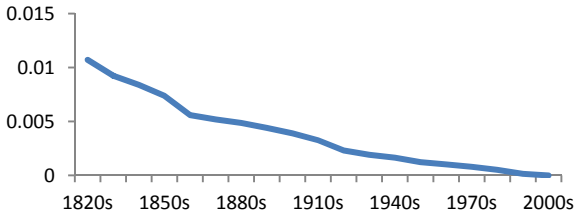

Figure 6 POS-based JS divergence between the last decade and other decades in COHA.

3.4 Word Lengths across Time

Word length is regarded as a key factor of readability. Figure 7 shows the length distributions of words in 1820s and 2000s (top and bottom graphs, respectively) based on Google 1-gram dataset.

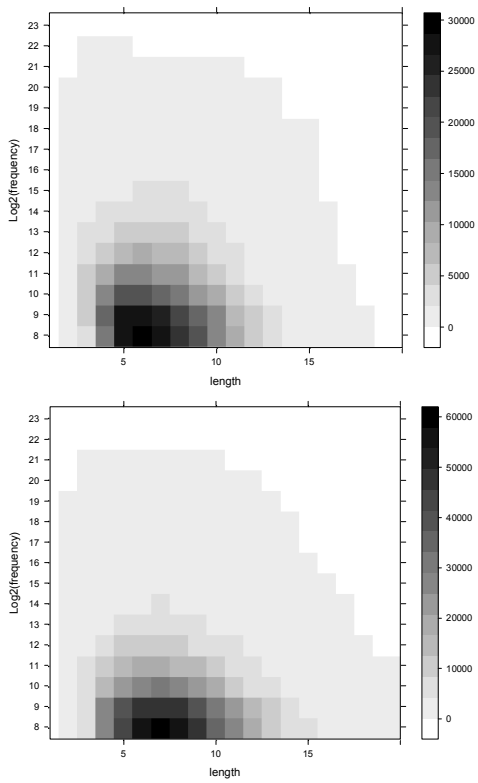

Figure 7 Distributions of words vs. their length and frequency in 1820s (top) and in 2000s (bottom) in Google 1-gram dataset.

The horizontal axis denotes the word length by characters. The vertical axis denotes the frequency of a word in a decade, while the color indicates the total number of unique terms that correspond to the given length and frequency. Comparing both graphs we can see that the plot for 2000s is wider and relatively flatter in terms of the numbers of unique words, meaning that, on relative basis, more long words are used now than before. For example, there is certain number of unique terms over 18 characters long in 2000s, while no such tokens in 1820s.

We note that many readability indexes such as Flesh-Reading Ease [5], Coleman-Liau [6] or Dale-Chall [2] include word length as a key factor, although, more complex readability measures were proposed (e.g., [1]). Based on the above results we think that decrease in readability experienced by current readers accessing older documents [6] could be due to other factors than word length. The decreased average length of words in past decades when compared to the more recent ones is actually the opposite force. Thus other factors such as longer sentence length [10], change in grammar structures, unknown wider background and missing context could influence the readability decrease of historical texts from the viewpoint of today's readers. To answer this question one should however perform user studies.

4. CONCLUSIONS

Language is a constantly evolving tool whose current state is the cumulative product of long and short-term evolution. In this paper we quantify various aspects of change in active English vocabulary over the last two centuries based on two large lexical corpora in order to better understand its impact on readability and accessibility of historical texts.

5. ACKNOWLEDGMENTS

This research was supported in part by: MEXT Grant-in-Aid for Scientific Research (#24240013) and for Young Scientists B (#22700096), and by the JST research promotion program Sakigake: "Analyzing Collective Memory and Developing Methods for Knowledge Extraction from Historical Documents".

6. REFERENCES

[1] K. Collins-Thompson and J. P. Callan. A Language Modeling Approach to Predicting Reading Difficulty. In *HLT-NAACL 2004*

[2] E. Dale and J.S. Chall. The Concept of Readability, *Elementary English*, 26(23), 1949

[3] A Garcia-Fernandez et al. When was it Written? Automatically Determining Publication Dates. In *SPIRE 2011*

[4] L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively Motivated Features for Readability Assessment. In *European Conference for Computational Linguistics, 2009*

[5] R. Flesch. A New Readability Yardstick, *Journal of Applied Psychology*, 1948, 32(3), pp. 221-233.

[6] A. Jatowt and K. Tanaka, Longitudinal Analysis of Historical Texts' Readability, In *JCDL 2012*, pp. 353-354, 2012

[7] N. Kanhabua and K. Nørvåg. Using Temporal Language Models for Document Dating, Machine learning and knowledge discovery in databases, Springer LNCS, 2009, Vol. 5782, 738-741

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

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

[10] L.A. Sherman. Analytics of Literature: *A Manual for the Objective Study of English Prose and Poetry*, Boston-Ginn, 1893.