Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3087–3095
Marseille, 11–16 May 2020

WordWars: A Dataset to Examine the Natural Selection of Words

Saif M. Mohammad
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca

Abstract

There is a growing body of work on how word meaning changes over time: *mutation*. In contrast, there is very little work on how different words compete to represent the same meaning, and how the degree of success of words in that competition changes over time: *natural selection*. We present a new dataset, *WordWars*, with historical frequency data from the early 1800s to the early 2000s for monosemous English words in over 5000 synsets. We explore three broad questions with the dataset: (1) what is the degree to which predominant words in these synsets have changed, (2) how do prominent word features such as frequency, length, and concreteness impact natural selection, and (3) what are the differences between the predominant words of the 2000s and the predominant words of early 1800s. We show that close to one third of the synsets undergo a change in the predominant word in this time period. Manual annotation of these pairs shows that about 15% of these are orthographic variations, 25% involve affix changes, and 60% have completely different roots. We find that frequency, length, and concreteness all impact natural selection, albeit in different ways.

Keywords: Natural selection, lexical semantics, words, evolution, word length, frequency, concreteness

1. Introduction

We rely on words to articulate our thoughts. The goal is often not just to be precise, but also to be clear and compelling.[1] Since multiple words can represent the same meaning (for example, *allowable* and *permissible*), we often make decisions on which words to use from sets of known synonyms.[2] For example, in the current day, it more common to say:

talking while chewing is not permissible

rather than *allowable*. Collectively, as speakers of a language we choose some words more frequently than others, to represent a meaning. Occasionally new words emerge, and every now and then some words fall out of favour so much that they may be considered *extinct*.

Thus it is not surprising that prominent thinkers of the past, including Charles Darwin, have argued that just as there is a natural selection in the plant and animal species over time, there is a natural selection of words over time (Darwin, 1968; Darwin, 2003). In this analogy, *word type* maps to a species, and *word frequency* maps to the species population. Thus the creation and use of a new word is akin to the birth of a new species, the wide-spread use of a word is akin to the thriving of a species, and the complete lack of use of a word is akin to the extinction of a species.

Over the last few years, there has been a spurt of research on historical language change, especially on how the meaning of a word has changed over time (Mihalcea and Nastase, 2012; Hamilton et al., 2016a; Bamler and Mandt, 2017; Zimmermann, 2019; Jawahar and Seddah, 2019)—this might be considered as a mutation. Much of this work is a direct result of the availability of resources such as the Google Books Ngrams Corpus (GBNC) (Google, 2012). However, there is very little computational work on the natural selection of words; that is, how words compete to represent a meaning, and how, over time, some words become more successful, whereas others become less successful.

In this paper, we present a large dataset to foster computational studies on the natural selection of English words. Specifically, the dataset includes sets of synonymous monosemous words and their frequencies in English books across 200 years.[3] We focus on monosemous words because readily accessible large amounts of text is not sense-annotated and thus it is difficult to track frequencies of individual word senses. However, working with polysemous words remains interesting future work.

The words are taken from the English WordNet 3.1 (Fellbaum, 1998). Historical frequencies are obtained from the English GBNC (Google, 2012). Since the dataset includes sets of words that compete with each other to represent meanings, we refer to it as the *WordWars* dataset. For any given span of time, we will refer to the most frequent (monosemous) word in a synset as the *predominant* or *winner* word. WordWars includes the predominant word and frequency information for 5,062 synonym sets from 1800 to 2009. To facilitate easy exploration of the WordWars dataset, we also created an online interactive visualization to illustrate individual battles within each of the synsets.

We begin with an overview of the related work in Section 2 Sections 3 describes the WordWars dataset: how we created it (§3.1) and how we visualize it (§3.2). In Section 4, we explore three questions on the natural selection of monosemous words, using the WordWars dataset:

1. What proportion of predominant words from early 1800s are displaced by other words in the 2000s (both overall and across select sub-classes such as part of speech)? (§4.1)

2. Which word features have a significant impact on the natural selection of words, i.e., on whether a word that thrived in the early 1800s will be displaced by another in the early 2000s? We explore three word features of particular importance in linguistic and psychological

[1] For example, when writing this introduction.

[2] One can argue that no two words are exactly the same in meaning, and thus there are no true exact synonyms, but only near-synonyms. However, for the sake of brevity, in this paper we will use the term *synonyms*, rather than *near-synonyms*.

[3] Monosemous: words that have exactly one meaning.

studies: frequency (in early 1800s), length, and concreteness. (§4.2)

3. What are some of the notable differences between the predominant words of early 1800s and the predominant words of early 2000s, especially in terms of orthography, affixation, and word length? (§4.3)

WordWars can be used to explore a number of other research questions as well, including those listed in Section 5 We leave that for future work.

All of the data associated with this project, as well as the interactive visualizations, are made freely available through the project webpage.[4]

2. Related Work

In his seminal work, *The Descent of Man*, Charles Darwin recognized that evolution is not just a feature of biological entities but also of words:

> *the survival or preservation of certain favoured words in the struggle for existence is natural selection*
>
> *(Darwin, 2003)*

Interestingly, the term *evolution* and associated concepts were used to describe cultural and language changes long before Darwin applied them to living organisms (van Wyhe, 2005). Language evolution has since been of substantial interest in a number of fields, especially linguistics, cultural studies, phonosemantics, psychology, and evolutionary studies. See survey articles for recent research on language change (Gray et al., 2007; Pagel, 2009; Mesoudi, 2011). Below we identify some of the most relevant work to the current study.

George Zipf famously stated that:

> *the magnitude* [length] *of words tends to stand in an inverse relationship to the number of occurrences* *Zipf (1949)*

Thus shorter words tend to be more frequent than longer words. He attributed this to the human tendency to preserve effort.

Concreteness measures the degree to which a word refers to an entity perceivable by the senses. Some studies argue that concrete words are easier to remember because they trigger not just the verbal codes of memory but also the perceptual ones (Paivio, 1971).

Similar to concreteness, it has been shown that high frequency terms and longer words are more easy to recall in lexical decision tasks (Meyer and Schvaneveldt, 1971; Schvaneveldt and Meyer, 1973; Barber et al., 2013; Keuleers et al., 2010).

Frequency, word length, and concreteness have also been shown to play an important role in child language acquisition. Children learn the high-frequency, concrete, and shorter words first (Goulden et al., 1990). In this work we examine the impact of frequency, word length, and concreteness on the evolution of words.

[4] http://saifmohammad.com/WebPages/wordwars.html

Bolinger (1953) claimed that words that sound similar to near-synonyms have a greater chance of being more frequently used. Magnus (2001) shows that some phonemes tend to be used in a restricted semantic space, i.e., they occur in words with similar meanings. For example, labial consonants like /b/ tend to convey a sense of roundness, as in, *ball, bell, boat, blob, blotch, bun,* and *bulb*. Cuskley et al. (2014) shows that even though some forces push verbs to become regular, other forces regenerate and maintain irregularity in language. We do not examine these ideas here, but WordWars can be of benefit for that work as well.

There is growing interest in Computational Linguistics research in studying language change. A large majority of that work examines how the meaning of a word changes over time (Mihalcea and Nastase, 2012; Hamilton et al., 2016a; Hamilton et al., 2016b; Bamler and Mandt, 2017; Dubossarsky et al., 2017; Rudolph and Blei, 2018; Zimmermann, 2019; Jawahar and Seddah, 2019). For example, *gay*, *wicked*, and *nice* mean very different things now than what they did in the past. Essentially, that work examines how a word gains new senses, and how some senses of a word may become deprecated. In contrast, here we examine how different words compete to represent the same meaning, and how the degree of success of words in that competition changes over time.

WordWars can also be used to develop various machine learning systems, including those that aim to predict the progression of word frequencies over time. See Turney and Mohammad (2019) for a supervised learning algorithm that is able to predict the future leader of a synset—the word in the synset that will have the highest frequency.

3. WordWars: Historical Frequencies of Sets of Competing Words

We begin with how we created WordWars (§3.1), followed by how we visualize it (§3.2).

3.1. Creating WordWars

Our goal was to compile a dataset of groups of words that (roughly) mean the same thing, along with their frequencies of usage over time. Since some words have more than one sense and it is difficult to determine the intended sense automatically in text, we chose to focus on monosemous words (words with exactly one sense or meaning). We use WordNet 3.1 (Fellbaum, 1998) to obtain sense information and the Google Books Ngrams Corpus 2012 version (GBNC) (Google, 2012) to obtain historical frequency information. The precise steps we used to create the dataset are listed below:

1. *Identify words that are monosemous using WordNet 3.1.* Words that occur in only one synset were considered monosemous (e.g.: *aptly*, whereas *madly*, which occurs in multiple synsets was ignored). Words that occur in the same WordNet synset are considered synonymous (e.g., *aptly, ably, competently* and *capably*). Table 1 rows a to d show the number of words and synsets in WordNet 3.1 (overall and monosemous).

a. # word types:	207,234
b. # synsets:	117,790
c. # monosemous word types:	131,291
d. # synsets with at least two monosemous words:	34,396
e. # monosemous word types with GBNC frequencies:	47,981
f. # synsets with at least one monosemous word and GBNC frequencies:	37,941
g. # battle synsets	7,665
adjective synsets:	1,560
adverb synsets:	464
noun synsets:	4,572
verb synsets:	1,069
h. # monosemous word types in battle synsets	17,705

Table 1: Number of words and synsets in WordNet 3.1 for which GBNC frequencies are available.

2. *Obtain historical unigram frequencies for the monosemous words from the GBNC.* Table 1 row e shows the number of WordNet monosemous words for which historical frequency information is available in GBNC. To smooth frequencies, we aggregate frequencies across ten-year spans. Thus the entry for a word w and 1809, includes the total number of times w occurred in books from 1800 up to and including 1809; the entry for 1810, includes the total number of times w occurred in books from 1801 up to and including 1810; and so on. Each word has frequency scores for all the ten-year spans from 1809 to 2009.[5] For brevity, we will refer to a span by its final year, i.e., the *1800–1809 span* will be referred to as the *1809 span*. (File *monosemous-word-historical-frequencies* has these tab-separated columns: word, ten-year span, frequency.)
3. *Identify battle synsets*. Determine synonym sets from WordNet 3.1 that include at least two monosemous words with unigram frequency information in the GBNC. These are the synsets where we examine the competition between the monosemous words to represent meanings. We will refer to them as the *battle synsets*. Rows g and h show the number of battle synsets and the number of monosemous words in the battle synsets, respectively.
4. *Determine predominant/winner words*. For a given span, we will refer to the most frequent (monosemous) word in a synset as the *predominant word* or *winner word*. We record the information of these words in the *winners file*, which has these columns: synset, winner word in the 1809 span, and winner word in the 2009 span. (File name: winners-1809-2009) A particularly useful subset of the WordWars dataset is the set of synsets where the winner word of the 2009 span is different from the winner word of the 1809 span. We will refer to these synsets as *change-of-winner synsets* and the pairs of winners as the *1809winner–2009winner pairs*. (File name: change-of-winner-1809-2009)
5. *Store contexts of words in battle synsets.* We extract all the 5-grams in the GBNC that include any of the words in the battle synsets. We do not make use of the 5-grams for analysis in this paper, but they provide valuable contexts for the competing words, and thus of potential use in future work. We do include the 5-grams in the interactive visualization so users can peruse the contexts of words at different stages of the word's life cycle (early adoption, when they become dominant, etc.).

We refer to the collection of files described above as the *WordWars Dataset*. We repeated these steps for two other periods of time: 1809 to 1909 and 1909 to 2009. An analysis of these additional datasets helps break down how language changed in each of the two corresponding centuries. We also created a version of the WordWars dataset from the GBNC Fiction subset. Although, the rest of the paper presents analyses only on the full dataset, experiments with the Fiction subset obtained similar results.[6]

3.2. Visualizing WordWars

The Google Ngram Viewer is an online interactive visualization where users can enter one or more words, and the system renders a graph of their frequencies over time (based on GBNC data).[7] One can enter the words from the change-of-winner synsets that WordWars provides to visualize the frequencies. However, entering words from hundreds of synsets can become tedious. To facilitate the examination of the competing words, we created an online interactive visualization to illustrate individual battles within each of the WordWars synsets. Figure 1 shows one of the sub-visualizations.

The user can view individual 1809 and 2009 winner pairs on the left. With a single click, one can choose to explore only the change-of-winner pairs. (The image in Figure 1 is after this click.) Clicking on any of the word pairs shows the historical relative frequencies of all the words in the corresponding synset in the graph on the right, where *relative frequency* is the ratio of the frequency of the word in a year to the total frequency of all words in that year. For example, in Figure 1, one can see the relative frequencies associated with the *ably–aptly* change-of-winner pair, which includes frequencies not just for *ably* and *aptly* but also of other monosemous words in that synset (*competently* and *capably*). From the graph, the user can also note that the winner changed in 1956 (hovering over the graph shows the year and precise relative frequency information). Using just the down (or up) arrow key, one can cycle through a large number of change-of-winner synset quickly and easily. Figures 2 and 3 show examples where one can see both relative frequencies and raw frequencies, of competing sets of words, from 1809 to 2009.

[5] Much fewer books were published prior to 1800.

[6] Some recent work has criticized the use of the full GBNC arguing that it includes information from a large number of technical publications (Pechenick et al., 2015). However, since this work focuses only on monosemous words and compares word frequencies only within a synset, we do not expect material from technical publications to impact the battles in non-technical synsets.

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

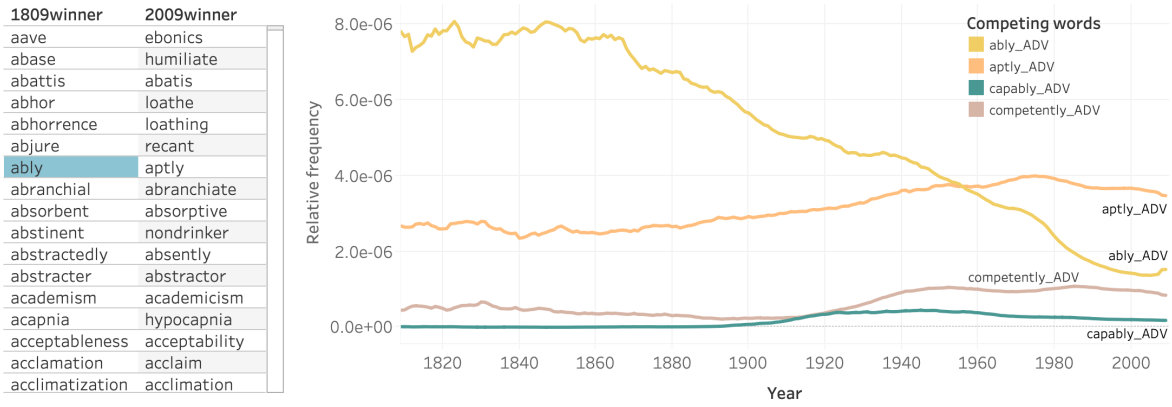

Figure 1: A subvisualization showing the historical relative frequencies of the words in a battle synset.

	span 1:	1809	1809	1909
	span 2:	2009	1909	2009
a. # monosemous word types with frequency (freq) > 0 in span 1:		26,264	26,264	41,279
b. # monosemous new words (freq $= 0$ in span 1 but freq > 0 in span 2):		21,698	15,249	6,694
c. # span 1 monosemous words that died out by span 2 (freq $= 0$ in span 2):		6	234	17
d. # monosemous word types with freq > 0 in span 2:		47,956	41,279	47,956
e. # battle synsets (synsets with more than one monosemous word with freq > 0 in span 2 and at least one monosemous word with freq > 0 in span 1):		5,062	4,747	6,708
f. # monosemous words in battle synsets:		11,896	11,028	15,498
g. # times the predominant word in span 2 $\neq$ the predominant word in span 1:		1,607	1,086	1,621

Table 2: Statistics from WordWars for three pairs of spans: 1809–2009, 1809–1909, and 1909–2009.

Other visualization options (not shown here), allow one to choose: (1) the base corpus (GBNC Full or Fiction), (2) to view specific kinds of change-of-winner pairs (orthographic variants, same root different affixation, or different root), (3) particular parts of speech (noun, verbs, adjectives, or adverbs), and (4) whether the synset includes a new word (a word that occurs in GBNC in any of years after 1809, but not in the 1809 span). One can also click on the desired ten-year span to see the contexts (5-grams) in which the word was used. This can help better understand for example, the contexts in which a word was used at different stages of its life cycle: for example, when people first started to use it, when its use increased or decreased markedly, when its use surpassed the previous predominant word, when it was close to extinction, etc. A slider allows users to select ranges of years such that the system will show only those change-of-winner synsets for which the usage of the 2009 winner surpassed the usage of the 1809 winner in the selected time period. We hope that the visualization will encourage further studies on word evolution.

4. Examining the Natural Selection of Words with WordWars

We now present the use of WordWars to better understand three broad questions associated with the natural selection of words: (1) the rate of change of predominant words (§4.1), (2) importance of three prominent word features in determining whether an 1809 winner will be displaced in 2009 (§4.2), and (3) differences between the 1809 and 2009 winners (§4.3).

4.1. Rate of Change of Predominant Words and the Impact of New Words

Table 2 provides a number of statistics from the WordWars dataset for three pairs of spans: 1809–2009, 1809–1909, and 1909–2009. Observe that only about 26,000 of the 47,981 words that occurred in the 2009 span, also occurred in the 1809 span. This is not surprising since the number of books in the Google Books Corpus from 2000 to 2009 is orders of magnitude larger than the number of books from 1800 to 1809. Interestingly though, more of the new words were first seen between 1809 and 1909 ($\sim$15K) than between 1909 and 2009 ($\sim$7K).

Note that the number of battle synsets for the 1909–2009 file (6,708) is markedly higher than the number of battle synsets for the 1809–2009 file (5,062). This is because of the higher vocabulary size in 1909, which leads to more synsets from that time having a non-zero number of monosemous words.

Observe that in 32% of the 2009 span battle synsets, the predominant word is different from the predominant word in in the 1809 span (1,602 of the 5,062). The percentage is lower for the 1809–1909 span pair (29%) and for the 1909–2009 span pair (24%). This makes sense because there is a greater chance for the dominant word to change over a longer period of time as compared to over a shorter period of time. It is interesting to note that the percentage of change is higher in the 1809–1909 span pair than in the 1909–2009 pair. Perhaps this is correlated with the fact that the former period saw a larger number of new words, than the latter period.

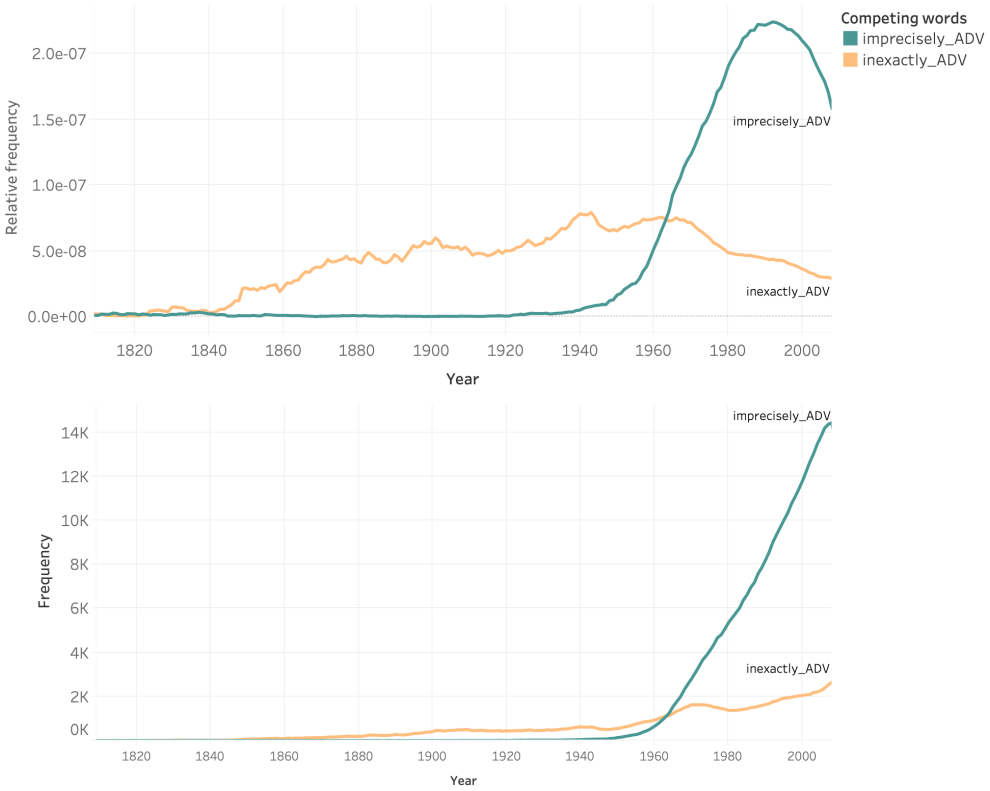

Figure 2: Relative frequencies and raw frequencies of the words *imprecisely* and *inexactly* from 1809 to 2009.

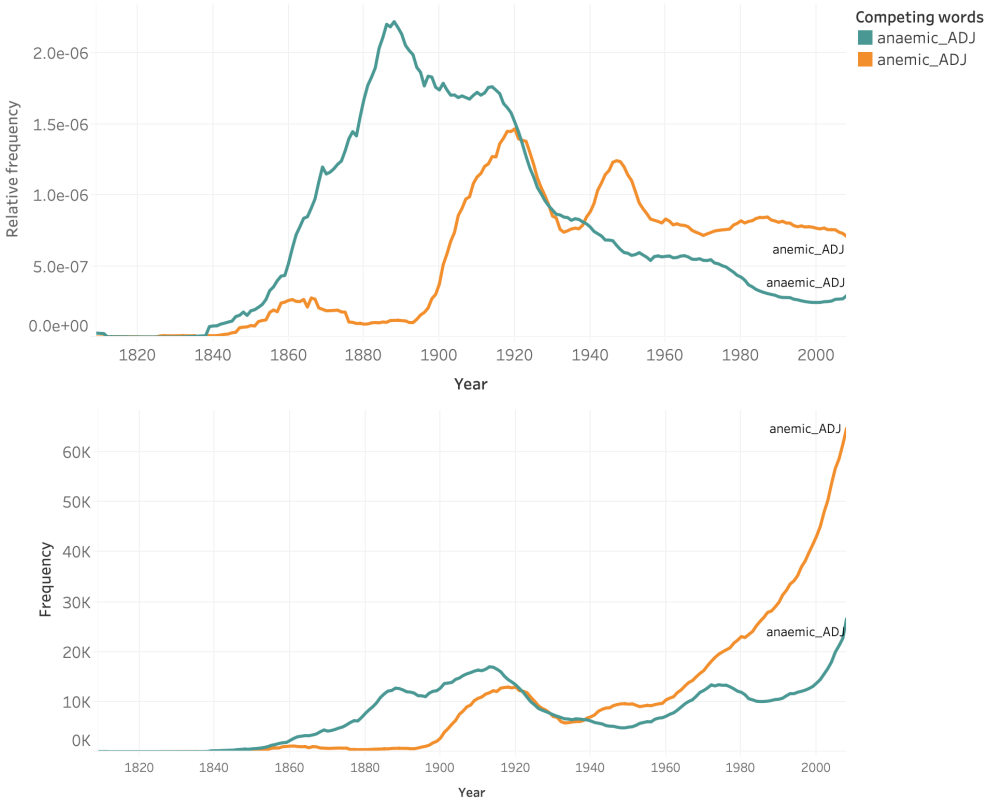

Figure 3: Relative frequencies and raw frequencies of orthographic variants *anaemic* and *anemic* from 1809 to 2009.

span 1:	1809	1809	1909
span 2:	2009	1909	2009
# battle synsets: overall	5,062	4,747	6,708
: with new words	2,304	1,972	1,046
# new words in battle synsets	2,774	2,274	1,199
# new words that won battles	593	325	271

Table 3: Statistics on new words in WordWars.

span 1:	1809	1809	1909
span 2:	2009	1909	2009
adjectives	216	160	202
adverbs	115	66	92
nouns	1,169	791	1,209
verbs	107	69	118
all	1,607	1,086	1,621

Table 4: Number of change-of-winner pairs in various span pairs, by part of speech.

Table 3 shows the number of new words in different span pairs, the number of synsets they are involved in, as well as the number synsets where the new words eventually became the predominant words. Observe that for the 1809–2009 span pair, 45% (2304/5062) of the battle synsets included at least one new word (a word that did not occur in the ten years from 1800 to 1809). Further, in 25% of the synsets (593 of the 2304 synsets), the new word eventually became the predominant word.

Table 4 shows a break down of the number of predominant word changes by part of speech. Observe that there are a far greater number of changes in noun synsets than any other part of speech. However, the numbers are roughly proportional to the number of battle synsets for each part of speech.

4.2. The Role of Frequency, Length, and Concreteness in Determining Whether an 1809 Winner will be Displaced in 2009

A compelling question in the understanding of language change is why some words that once thrived (in terms of speaker preference) are displaced by others over time. We now use the WordWars dataset to examine this question. Some words that won the dominance battles in the 1809 span, also won the battles in the 2009 span, whereas others did not. We will refer to the former set of words as the *1809win2009win* words and the latter as the *1809win2009loss* words. We determine how the two sets of words compare in terms of three word features prominently studied in linguistics, cognitive science, and psychology: (a) frequency (in the 1809 span), (b) length, and (c) concreteness. We use the concreteness ratings for about 40K English lemmas compiled by Brysbaert et al. (2014).[8] The ratings are real values between 1 and 5.

Since the distributions of word frequencies tend to not have a normal distribution (as is also the case in our dataset), we use the two sample Kolmogorov–Smirnov test (K–S test) to determine whether the 1809 span frequencies of the 1809win2009win words are significantly different from the 1809 span frequencies of the 1809win2009loss words (Massey Jr, 1951). Since word length and concreteness ratings have normal distributions, we used the Student's t test to determine whether the word lengths and concreteness ratings of the 1809win2009win words and the 1809win2009loss words are significantly different. Table 5 shows the sample sizes and means of the relevant distributions.[9]

The significance tests show that low frequency 1809 winners are significantly more likely to be displaced by other words of their synsets, compared to high frequency 1809 winners ($p < .001$). This makes sense as a competitor does not need to gain an overly large frequency to replace the 1809 winner word. Nonetheless, it also shows that in low-frequency synsets the change in relative frequencies of its members can be more dramatic than in high-frequency synsets. The length of a word does not have a significant impact on whether it is likely to be replaced or not ($p > .05$). The concreteness of the word has a significant impact on whether it will be replaced: low concreteness (i.e., abstract) 1809 winners are more likely to be displaced than high concreteness 1809 winners ($p < .001$). It is worth exploring, why: Is this because abstract words by their nature tend to be displaced more often than concrete words? Or, is this because there has been a much larger proliferation of abstract words (than concrete words) in the last two hundred years? Or, is there some other cause? We leave that for future work.

4.3. Differences between the 1809 Winners and the 2009 Winners

Here, we examine the differences between the predominant words of 1809 and the predominant words of 2009. We do so by manually annotating the 1809 and 2009 winner pairs into three types: orthographic (spelling) variants, same root but different affixation, and completely different roots.[10] We discarded 66 of the 1607 pairs from further analysis as they did not fall into one of the three pair types (e.g., word and its abbreviation).

Table 6 shows numbers for each type and examples of sub-types. Observe that close to 60% of the pairs have different roots, about 25% have the same root but different affixation, and about 15% are orthographic variants. The individual types can be further examined to better understand changes in orthography, affixation, and how a completely different word can come to dominate a synset. In the subsections below we explore (1) changes in affixation, and (2) difference in lengths of the 1809 winner and the 2009 winner, which is uniquely related to all three pair types.

[8] http://crr.ugent.be/archives/1330

[9] The concreteness experiment involves a subset of the 1809 winners for which concreteness ratings are available. Therefore, the smaller sample size (N concreteness).

[10] The annotation was done by the author. We explored options of automatically classifying word pairs into the three types using available stemmers, etymology resources, etc., but found the results to be lacking both in coverage and accuracy. With the manually annotated dataset, we can be more confident about the conclusions, and further, we expect the annotations to be a useful resource to others using or studying the WordWars dataset.

word sets	Samples N	Mean 1809 span frequency	Mean length	Samples N concreteness	Mean concreteness
a. 1809win2009loss	1,601	1,233	8.807	484	2.834
b. 1809win2009win	3,430	3,216*	8.468	1,899	3.023*

Table 5: Mean 1809 span frequencies, mean lengths, and mean concreteness of (a) the words that won both in 1809 and 2009, (b) the words that won in 1809 but lost in 2009. A * next to a mean of a distribution in row (b) indicates that the distribution is statistically significantly different from the corresponding distribution of row (a).

Word pair type: #pairs	Example pair
Orthographic variants: 271	
letter deleted	*anaemic–anemic*
letter added	*wilfully–willfully*
letter replaced	*geodetic–geodesic*
letter transposed	*litre–liter*
Same root, diff. affix: 365	
different prefix	*maltreat–mistreat*
different suffix	s͡parseness–sparsity
prefix deletion	*benumb–numb*
suffix deletion	*burglarize–burgle*
Different roots: 905	*filch–pilfer, garb–attire*

Table 6: Types of changes in winner words.

Affix Change	# Flips	Example pair
suffix change		
ical–ic	32	*rhythmical–rhythmic*
ableness–ability	18	*adorability–adorableness*
re–er	8	*meagre–meager*
y–e	7	*competency–competence*
ous–ic	7	*nitrous–nitric*
ing–NONE	7	*yawning–yawn*
prefix change		
un–non	6	*unrenewable–nonrenewable*
un–in	6	*unadvisable–inadvisable*
ana–an	3	*anaesthetic–anesthetic*

Table 7: Most frequent affix-pair changes from the 1809 winners to the 2009 winner.

4.3.1. Changes in Affixation

Among the 365 same root change-of-winner pairs, we observe that suffix changes are markedly more numerous than prefix changes. Table 7 shows the most common affix and prefix changes. Table 8 shows the top five and bottom five suffixes by *net gain*—number of additional 2009 winner words that have the suffix. The "NONE" entry indicates that the 2009 winner was often a word resulting from the dropping of a suffix from the 1809 winner. The five entries in the middle of Table 8 correspond to common suffixes that were similarly dominant both in the 1809 and the 2009 spans. There was markedly less change in the prefixes of winners in the 1809 span and the 2009 span, with the notable exception of *non* and *re* (which have become markedly more frequent in the 2009 span) and *un* and *ana* (which have become markedly less frequent in the 2009 span).[11] The full lists of suffixes and prefixes ordered by net gain are available on the project page.

[11]Table with examples not shown due to space constraints.

Suffix	Y1 wins	Y2 wins	Net Gain
Top five by Net Gain			
ic	106	147	41
NONE	22	53	31
ability	33	51	18
or	58	67	9
ity	133	139	6
High frequency prefixes with close to 0 Net Gain			
ization	30	30	0
able	34	34	0
ate	56	56	0
ator	23	23	0
ness	107	107	0
Bottom five by Net Gain			
ous	47	34	-13
y	609	595	-14
ableness	20	3	-17
al	143	121	-22
ical	69	40	-29

Table 8: Suffixes in the 1809 and the 2009 winners.

4.3.2. Lengths of New Winners

If the principle of least effort for speakers, championed by Guillaume Ferrero and George Zipf (Ferrero, 1894; Zipf, 1949), is one of the forces in language evolution, then new winners are expected to be shorter in length than the earlier winners (a mechanism to reduce speaker effort). We test this hypothesis by comparing the lengths of the 2009 winners and the 1809 winners they have displaced. A one-tailed t test shows that the mean length of the 2009 winners is statistically significantly smaller than the mean length of the words they have displaced ($p < .001$). Table 9 shows the percentage of 2009 winners that are shorter, longer, and the same length as the 1809 winners they have displaced. If the change in length was random, then these percentages would each be 33%. However, row a. shows that overall, there is a markedly greater tendency for a word to be displaced by a shorter word. The rows b, c, and d show the results for three word-pair types. Notice that the tendency towards shorter words is most pronounced in same-root (affix-change) pairs. On the other hand, note that for noun and verb different-root pairs, it is more common for the 2009 winner to be longer than the word it displaces. The tendency for longer words is especially strong in adverb, noun, and verb orthographic variations, yet, adjectival orthographic variations lead markedly to shorter words. Thus, even though overall there is a tendency towards shorter words, within particular types and parts of speech the tendency can be towards longer words. This raises new questions as to why these differences exist.

Pair Type	# terms	Y2 winner word is shorter	longer	same
a. all	1,541	**41.1**	36.7	22.2
b. orthog. var.	271	**38.4**	23.6	**38.0**
adj	30	**63.3**	23.3	13.3
adv	8	25.0	25.0	**50.0**
noun	216	36.6	23.1	**40.3**
verb	17	23.5	29.4	**47.1**
c. same roots	365	**47.4**	34.8	17.8
adj	135	**49.6**	31.1	19.3
adv	9	**66.7**	33.3	0.0
noun	207	**45.9**	37.2	16.9
verb	14	**35.7**	**35.7**	28.6
d. different roots	905	39.2	**41.6**	19.2
adj	46	**47.8**	41.3	10.9
adv	100	**49.0**	34.0	17.0
noun	681	37.5	**42.7**	19.8
verb	78	37.2	**41.0**	21.8

Table 9: Percentage of 2009 winner words that are shorter, longer, or of the same length as the 1809 winners they displaced. Highest scores in each row are shown in bold.

5. Other Research Questions that can be Studied with the WordWars Dataset

The WordWars dataset can be used to study various aspects of language change, including but not limited to:

- To study the rate at which language is changing in different time periods within 1809–2009. For example, to identify short periods of time that incurred a large number of changes in the dominant word.
- To study word attributes (beyond those examined here) that are correlated with greater success (higher relative frequencies).
- In this work we explored the tendency of predominant words to be displaced. However, WordWars can be used more generally to study changes in usage frequency of a word relative to its synonyms.
- When plotting the relative frequencies of words across time, we observe various shapes. These shapes can be categorized into different kinds to study the which shapes are more common, and what factors influence the shape.
- To develop supervised machine learning models that are able to predict the future winner of a synset. We have already done some initial work in that regard (Turney and Mohammad, 2019).
- The WordWars dataset along with the Google 5-grams dataset can be used to study word usages immediately after birth, when the word becomes the most dominant word in the synset, when another word displaces it and it is no longer dominant, when it is close to extinction, etc. Such analyses can yield socio-cultural clues into the use of words.

WordWars is made freely available for research.

6. Limitations

The study of the natural selection of words is challenging for several reasons, including the lack of linguistic resources for words from earlier centuries, difficulty of tracking the frequencies of different senses of a word, and the occasional changes in the meanings of words over time. We have taken some steps to alleviate the impact of these challenges, including: focusing on monosemous words, using resources like WordNet that have substantial coverage (even of deprecated words), and limiting the analysis to only 200 years (the proportion of words whose meaning has changed dramatically in just 200 years is expected to be small). Nonetheless, it should be noted that these factors may introduce certain biases.

7. Conclusions

We presented a new dataset, *WordWars*, with historical frequency information from the early 1800s to the early 2000s for monosemous words in over 5000 synsets. We used the dataset to analyze several research questions about the natural selection of words in that time period.

We showed that close to one third of the synsets undergo a change in the predominant word in this time period. Manual annotation of these pairs shows that about 15% of these are orthographic variations, 25% involve affix changes, and 60% have completely different roots. We showed that even though length does not determine whether a predominant word will be displaced, if displaced, the new predominant word tends to be shorter. Notable exceptions for certain word-pair types and parts of speech are also identified. We showed that concreteness and frequency are significant factors in determining whether an 1800s predominant word will be displaced by the early 2000s.

We also presented a ranking of affixes by the extent to which they have become more popular in this time period. We showed that the suffixes *ic* and *ability* and prefixes *non* and *re* gained the most popularity whereas suffixes *al* and *ical* and prefixes *un* and *ana* dropped the most in popularity.

Finally, we identified several research questions that can be studied with the WordWars dataset. The dataset and the interactive visualization are made freely available.[12] We hope that these will foster further research on the natural selection of words.

Future work will explore extending this work beyond monosemous words to all words (which will require identifying ways to disambiguate word senses with sufficient accuracy) and analyzing data in additional languages to determine how patterns of natural selection may vary. We are also interested in exploring the complex inter-relationship between the change of meaning of a word over time and its capacity to compete with other words for favor in usage.

Acknowledgments

Many thanks to Peter Turney for the encouragement to pursue this project and helpful discussions.

[12] http://saifmohammad.com/WebPages/wordwars.html

8. Bibliographical References

Bamler, R. and Mandt, S. (2017). Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org.

Barber, H. A., Otten, L. J., Kousta, S.-T., and Vigliocco, G. (2013). Concreteness in word processing: Erp and behavioral effects in a lexical decision task. *Brain and language*, 125(1):47–53.

Bolinger, D. (1953). The life and death of words. *The American Scholar*, 22(3):323–335.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., and Tria, F. (2014). Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of english. *PLOS ONE*, 9(8).

Darwin, C. (1968). *On the Origin of Species*. Penguin, London, UK. Original edition, 1859.

Darwin, C. (2003). *The Descent of Man*. Gibson Square, London, UK. Original edition, 1871.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1136–1145.

Christiane Fellbaum, editor. (1998). *Wordnet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ferrero, G. (1894). L'inertie mentale et la loi du moindre effort. *Revue Philosophique de la France et de l'Étranger*, 37:169–182.

Google. (2012). Google books ngram corpus.

Goulden, R., Nation, P., and Read, J. (1990). How large can a receptive vocabulary be? *Applied linguistics*, 11(4):341–363.

Gray, R., Greenhill, S., and Ross, R. (2007). The pleasures and perils of Darwinizing culture (with phylogenies). *Biological Theory*, 2(4):360–375.

Hamilton, W., Leskovec, J., and Jurafsky, D. (2016a). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

Jawahar, G. and Seddah, D. (2019). Contextualized diachronic word representations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*.

Keuleers, E., Diependaele, K., and Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, 1:174.

Magnus, M. (2001). *What's in a Word? Studies in Phonosemantics*. Ph.D. thesis, Norwegian University of Science and Technology, Trondheim, Norway.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Mesoudi, A. (2011). *Cultural Evolution*. University of Chicago Press, Chicago, IL.

Meyer, D. E. and Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90(2):227.

Mihalcea, R. and Nastase, V. (2012). Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 259–263.

Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, 10:405–415.

Paivio, A. (1971). *Imagery and verbal processes*. New York.

Pechenick, E., Danforth, C., and Dodds, P. (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*.

Rudolph, M. and Blei, D. (2018). Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee.

Schvaneveldt, R. W. and Meyer, D. E. (1973). Retrieval and comparison processes in semantic memory. *Attention and performance IV*, pages 395–409.

Turney, P. D. and Mohammad, S. M. (2019). The natural selection of words: Finding the features of fitness. *PloS one*, 14(1):e0211512.

van Wyhe, J. (2005). The descent of words: Evolutionary thinking 1780-1880. *Endeavour*, 29(3):94–100.

Zimmermann, R. (2019). Studying semantic chain shifts with word2vec: Food¿ meat¿ flesh. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 23–28.

Zipf, G. K. (1949). Human behavior and the principle of least effort.