The great temptation: What diachronic corpora do and do not reveal about social change

Martin Hilpert

Abstract

This paper examines the potential that large diachronic corpora hold for the study of social change. Resources such as the COHA or Google Books make it possible to detect historical shifts in the frequencies of linguistic elements, which can then be interpreted as reflections of ongoing developments in society. But how should this be done in practice? This paper addresses this question in two main parts. The first, theoretical part surveys a series of problems that need to be controlled for in analyses of diachronic textual data. The second part implements these ideas in a study of change in the English *make*-causative over the past 150 years. Examining the variables of animacy and verb semantics as well as distributional collocational evidence, the study explores whether the diminishing social value of interpersonal authority is reflected in changing patterns of language use.

1 Introduction

Over the past decade, the empirical foundations of historical corpus linguistics have been strengthened considerably. Amongst other new resources, large diachronic corpora such as the COHA (Davies 2012) or Google Books (Michel et al. 2010) offer an unprecedented amount of recent historical data. These resources are highly useful for corpus linguists who are interested in lexical, grammatical, or stylistic developments, but even beyond that, researchers from other academic disciplines have realized the potential that large diachronic corpora hold for their research questions. In the words of Michel et al. (2010: 176), interpreting frequencies from diachronic corpora as reflections of ongoing developments in the world can "provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology". It seems that the possibilities are endless, and that diachronic linguistic data indeed has the potential of revealing to us everything we have ever wanted to know about social change, be it cultural, technological, political, or indeed linguistic. The

data is freely available, all that needs to be done is to bring the evidence to light. It is this very tempting idea that the title of this paper alludes to. The title also foreshadows that things may not be as easy as they appear at first sight. But what exactly are those problems, and what are the benefits that we can realistically expect?

A good starting point for a discussion of these questions is a study that illustrates how diachronic corpus frequencies can be used to explore social change. Greenfield (2013) takes positive frequency trends of words such as *choose* and *get* in the Google Books corpora as evidence for increasing individualism and materialism in American culture. The study is anchored in a broader theory of social change and human development (Greenfield 2009), which links fundamental social trends such as urbanization and economization to a theoretical model that distinguishes the notions of *gesellschaft* (society) and *gemeinschaft* (community). Common features of a *gesellschaft* environment are individualism and competition, whereas the members of *gemeinschaft* environments are relatively more generous, charitable, and interdependent. Over time, American culture has been adopting more and more *gesellschaft*-adapted features. Greenfield (2013: 2) hypothesizes that this is mirrored in language use, and that trends in linguistic frequencies reflect social developments:

> The goal of the study was to demonstrate that, as the United States moved ever further in the gesellschaft direction, gesellschaft-adapted cultural features, as indexed by relevant words in the corpus of millions of American books analyzed by the Google Ngram Viewer, showed a quantitative increase, whereas gemeinschaft-adapted cultural features, indexed by relevant words in the same corpus, showed a quantitative decrease.

Among the words that Greenfield (2013: 5-7) investigates are the elements *give*, *obedience*, and *authority*, which reflect *gemeinschaft*-adapted features, and *get*, *choose*, and *individual*, which reflect *gesellschaft*-adapted features. Figure 1 shows relative frequency trends from the Google Ngram Viewer for these elements. The x-axis shows the historical period of the data; the y-axis shows relative frequency as the percentage of the respective element in the corpus for the relevant year.

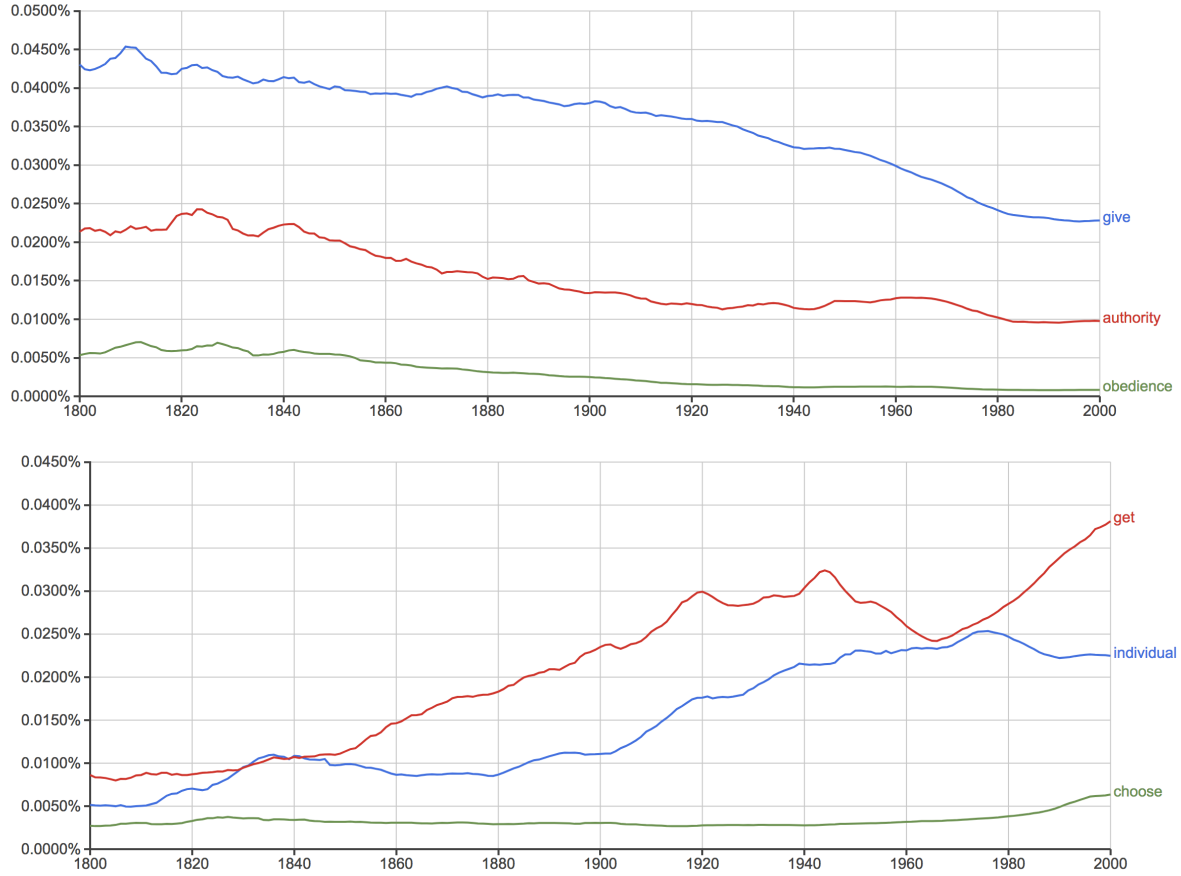

Figure 1: Relative frequency trends for six lexical elements in the Google Books corpora

It is readily apparent that *gemeinschaft*-adapted lexis is on the decline, while the reverse is true for *gesellschaft*-adapted words, which is of course in line with Greenfield's theoretical proposal. Her conclusion is that diachronic trends in word frequency do indeed mirror ongoing changes in society: "These findings signify that books as cultural products reflect human ecology. They also signify that cultural features can be indexed by word-use frequencies, which, in turn, reflect what is prioritized by a population" (Greenfield 2013: 8). Many members of the corpus-linguistic research community would be hesitant to generalize from a corpus, large as though it may be, to the general population. With the question of corpus representativeness (Biber 1993) at the heart of issues such as corpus compilation, register, variety, and comparisons between corpora, the limitations of individual corpora have been thoroughly discussed and recognized. Consequently, reactions to Greenfield's study from within linguistics have been skeptical. In a post on the *Language Log*, Liberman (2013) acknowledges that Greenfield's conclusions may actually be correct, but he argues

that the corpus analysis does not in fact show this: "I'm not arguing that her theory is wrong, or that the Google ngrams datasets don't contain supporting evidence. But it's going to take a much more careful and systematic analysis of the lexico-historical data to convince me." This comment raises the central question for this paper: How exactly can diachronic corpus data be analyzed in order to yield reliable insights about social change? How would a study have to be designed in order to be approved as "careful and systematic"? This paper will try to address this question in two parts. The first, theoretical part surveys the following five problems that need to be addressed in analyses of diachronic textual data.

1. Corpus frequencies are not always equivalent to frequencies of entities and events in the real world.
2. Corpus frequencies of polysemous words need to be broken down into sense-specific frequencies.
3. Correlations in large datasets may be spurious.
4. Comparisons of frequency trends in diachronic corpora require adequate statistical treatment.
5. It is not always easy to disentangle social change and linguistic change.

The second part implements these ideas in a study of change in the English *make*-causative (Kemmer 2001, Gilquin 2010). The study is inspired by an investigation by Verhagen (2000), who links diachronic changes in two Dutch causative constructions to a phenomenon of social change, namely the historical decline of interpersonal authority. As a grammaticalized expression of authority in English, the *make*-causative construction offers an appropriate test bed for the claims put forward by Greenfield (2013). Examples such as *She made the boys clean up their room* verbalize that a causer prompted a causee to perform a coerced action. If American culture becomes less authoritarian, as has been argued by Greenfield (2013), examples such as the one above should recede in favor of uses such as *The music made me smile*, which involve inanimate causers and non-coerced actions. Data from the COHA is retrieved to track the history of the *make*-causative. Examples are annotated in terms of several semantic parameters, including the animacy of causer and causee and the semantics of the verb that expresses the caused action. The analysis further draws on distributional collocational evidence by creating a semantic vector space for the verbs that

are found in the *make*-causative construction. This approach makes it possible to investigate which semantic classes of verbs have increased or decreased in their use with the *make*-causative construction. The results indicate that the construction has changed in a way that is consistent with the hypothesis of social change, but that differs in several respects from Greenfield's empirical observations, so that a common underlying cause for both developments is ultimately unlikely. A final discussion will explore whether social change is the only possible explanation for the trends in the data, or if processes such as subjectification and intersubjectification (Traugott 1989, 2010) can offer an alternative account.

2 Five pitfalls in the analysis of diachronic corpus data

2.1 Corpus frequencies (semasiological frequencies) are not always equivalent to frequencies of entities and events in the real world (onomasiological frequencies).

A basic assumption that underlies Greenfield's (2013) argument is that the frequencies of linguistic elements reflect their occurrence in the real world. If the word *obedience* is on the decline, this warrants the conclusion that speakers are less preoccupied with the idea of obedience, value it less, and are consequently less obedient in their behavior. This idea is in fact quite wide-spread, both within corpus linguistics and beyond. In the following quote, McEnery and Wilson (2001: 10) attribute the general point to Chomsky, and they indicate their agreement with it.

> As Chomsky himself stated so amusingly, the sentence *I live in New York* is fundamentally more likely than *I live in Dayton, Ohio* purely by virtue of the fact that there are more people likely to say the former than the latter.

Appealing as though the idea is, it is easy to find counterexamples. Table 1 shows frequencies from the NOW corpus (Davies 2013), specifically the frequencies of verbal elements that co-occur with the phrase *with a screwdriver*. The most frequent lexical verbs all refer to violent actions that are performed with a screwdriver. Thankfully, the frequencies

in Table 1 are in no way representative of the most frequent use of screwdrivers, which would be to fasten a screw. This goes to show that corpus frequencies and the frequencies of entities and events in the real world do not always match.

collocate	frequency		collocate	frequency
was	105		*been*	14
armed	80		*being*	13
stabbed	70		*forced*	13
threatened	42		*open*	13
attacked	33		*stab*	11
said	30		*found*	10
were	18		*can*	10
be	17		*is*	10

Table 1: Verbal collocate frequencies of "with a screwdriver" in the NOW corpus

A reasonable objection to this counterargument would be that the NOW corpus contains writing from news texts, and that ordinary events, like tightening a screw that has come loose, hardly merit any attention in a news article. While that point is well-taken, it raises the question why a noun such as *obedience* should be used more frequently if it describes a behavior that is seen as the norm rather than the exception. Should the basic assumption of an equivalence between corpus frequency and real-world occurrence frequency not be the same for all linguistic elements that are being investigated? Also, it is demonstrably not the case that news texts exclusively describe events that are unexpected. Elections, sports events, and political debates are regularly covered despite being fully expectable, and while the specific outcome of an event such as a football tournament may be unpredictable, most other aspects, including the participation of important players, the occurrence of offsides, and the location of semi-finals, are not. The NOW corpus further allows comparisons between events that are equally undeserving of media coverage. For example, the strings *doing my laundry* and *doing my taxes* occur with virtually identical frequencies, despite the fact that there is a distinct asymmetry in the real-world occurrence of the two events.

In order to come to terms with the problem at hand, it is in order to make a terminological distinction between two different types of frequency. The term *semasiological frequency* would correspond to the notion of text frequency as it is commonly used in corpus linguistics. Semasiological frequency captures how often speakers verbalize a given concept. The term *onomasiological frequency* captures how often speakers experience a given concept, without necessarily verbalizing it. Many mundane day-to-day activities, such as sitting down, opening a door, or doing one's laundry, are characterized by high onomasiological frequency but rather low semasiological frequency. For highly newsworthy events, the reverse is true. The bottom line for any analysis of corpus frequencies is that it has to be determined whether and how onomasiological and semasiological frequencies may differ for the elements that are being investigated. Neglect of this point may lead to erroneous conclusions, such as the idea that screwdrivers are highly dangerous objects, or that people in the United States do their laundry about as often as they do their taxes.

2.2 Corpus frequencies of polysemous words need to be broken down into sense-specific frequencies.

In a response to Greenfield (2013), Flach (2013) points out that frequency trends such as those depicted in Figure 1 cannot be taken at face value for elements that are highly polysemous. The verb *get* is a case in point, since it participates in many different grammatical patterns whose meanings do not instantiate the sense of receiving something. As is illustrated in (1), uses such as the *get*-passive or *get* as a copula are not usefully seen as reflections of a more individualistic culture.

(1)	a.	*get* 'receive'	I get more money now.
	b.	light verb uses	They get things done.
	c.	copula *get*	I'm getting ready.
	d.	the *get*-passive	She got hit.
	e.	particle verbs	We always get up early.
	f.	ditransitive *get*	She got him a present.

g. idiomatic uses I just don't get it.

If the verb *get* is to be used as evidence for the increasing establishment of selfishness, frequency counts would need to be adjusted in such a way that for example only instances with a following noun phrase would be included. That operationalization will still include false positives (*Can you get the door please*; They get things done; etc.) that would have to be excluded manually.

The observations made by Flach (2013) are in line with Liberman's call for an analysis that is careful and systematic. It is important to be precise about the patterns that are analyzed in order to minimize confounds. Flach's points also show that lexis as evidence for social change is potentially treacherous. Lexical items can undergo meaning change in relatively unpredictable ways, new items may replace old ones, and there is obviously the problem of which lexical items to select for analysis in the first place. In the light of this, there is probably a case to be made against relying only on lexical material in corpus-based studies of social change. Some of these problems are attenuated if grammatical constructions are used for that purpose. The reason for that is that grammatical elements are relatively more obligatory than lexical elements. Speakers cannot avoid them as easily as lexical items, and sometimes they cannot avoid them at all. The discussion in section 3 will return to this issue and present the example of a grammatical construction that may be both more useful and more reliable as an indicator of social change than lexical elements.

2.3 Correlations in large datasets may be spurious.

Within the corpus-linguistic community, it is generally accepted that both small and large corpora have their specific advantages. Whereas small corpora can cater to the needs of researchers who work with rich and highly precise annotations, large corpora are especially useful for data-intensive methodologies that require many types and tokens. It is clear that for some research questions, a given corpus can be too small. But can a corpus ever be too big? Can there be risks to using a large corpus? The answer to both questions is yes. Roberts and Winters (2013) point out a problem of large databases that is not commonly discussed

in corpus linguistics, namely that with increasing amounts of data, the likelihood of finding spurious correlations between elements in that data increases. Importantly, this tendency does not only hold for linguistic corpora, but for databases in general. A spurious correlation is a link between two variables that appears to be statistically solid, but which has no explanation in terms of any actual relation. Roberts and Winters (2013) offer several examples that involve language. For instance, countries with higher linguistic diversity show a higher frequency of lethal traffic accidents (2013: 2), and speakers who regularly take siestas are more likely to speak a language with relatively few verbal inflections (2013: 8). Based on these and other examples, Roberts and Winters argue that the noise-to-signal ratio in large databases has been substantially underestimated, and they urge readers to be more skeptical of correlational studies that do not implement the necessary controls. Coming back to Greenfield (2013) and the negative correlation of diachronic frequency trends for elements such as *give* and *get*, it is evident that large corpora will make it easy to find at least some lexical items that behave in accordance with a pre-existing hypothesis. In order to harness the statistical power that large corpora make available, a number of safeguards need to be implemented. The next section discusses one such safeguard that is specifically designed for the analysis of diachronic corpus data.

2.4 Comparisons of frequency trends in diachronic corpora require adequate statistical treatment.

While correlations between variables in large datasets are generally to be viewed with caution, this is even more so the case for correlating trends that can be extracted from diachronic data sources. Koplenig and Müller-Spitzer (2016) illustrate this point with an observation that, at first glance, seems to have a plausible explanation. Using data from the Google Books corpora (Michel et al. 2010), they measure lexical diversity, operationalized as type-token ratio, in different languages, including Spanish, German, French, as well as British and American English. With the exception of Chinese, all languages in the dataset show a correlation of type-token ratio with population size. While this correlation could be argued to be rooted in processes of language use, so that larger communities of speakers produce larger vocabularies, Koplenig and Müller-Spitzer (2016: 1) point out that there is an equally

strong correlation between lexical diversity in the investigated languages and global mean sea level, a variable that has no influence whatsoever on language use. This casts serious doubt on the conclusion that lexical diversity and population size are in fact related.

2.5 It is not always easy to disentangle social change and linguistic change.

3 Giving in to temptation: A case study of the English *make*-causative

4 Conclusions

References

Biber, Douglas. (1993). *Representativeness in corpus design*. Literary and Linguistic Computing, 8(4), 243-257.

Davies, Mark. (2012) The Corpus of Historical American English. Available online at https://corpus.byu.edu/coha/

Davies, Mark. (2013) Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. Available online at https://corpus.byu.edu/now/.

Flach, Susanne. (2013) Das Ego in der Sprache. Post on *SprachLog*. [http://www.sprachlog.de/2013/08/12/googleology-gets/] Accessed 31.12.2018.

Gilquin, Gaëtanelle. (2010). *Corpus, Cognition and Causative Constructions.* Amsterdam: John Benjamins.

Greenfield, Patricia M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological Science* 24, 1722-1731.

Kemmer, Suzanne. (2001). Causative Constructions and Cognitive Models: The English Make Causative. First Seoul International Conference on Discourse and Cognitive Linguistics: Perspectives for the 21st Century, 803-846. Seoul: Discourse and Cognitive Linguistics Society of Korea.

Koplenig, Alexander/Müller-Spitzer, Carolin (2016): Population size predicts lexical diversity, but so does the mean sea level – one problem in the analysis of temporal data. *PLOS ONE* 11(3).

Liberman, Mark. (2013). The culturomic psychology of urbanization. Post on *LanguageLog*. [http://languagelog.ldc.upenn.edu/nll/?p=5985] Accessed 31.12.2018.

Michel, Jean-Baptiste et al. 2010. Quantitative Analysis of Culture Using Millions of Digitized Books, *Science* 331: 176-182.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. Second edition. Edinburgh: Edinburgh University Press.

Roberts S, Winters J. Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits. *PLoS ONE* 8(8).

Szmrecsanyi, Benedikt (2016). About text frequencies in historical linguistics: disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory* 12(1): 153-171

Traugott, Elizabeth C. (1989). On the rise of epistemic meanings in English: An example of subjectification in semantic change. Language 57: 33-65.

Traugott, Elizabeth C. (2010). Revisiting subjectification and intersubjectification. In K. Davidse, L. Vandelanotte, and H. Cuyckens (eds.), *Subjectification, Intersubjectification and Grammaticalization*. Berlin: De Gruyter Mouton, 27-70.

Verhagen, Arie (2000), Interpreting Usage: Construing the history of Dutch causal verbs. In: Barlow M., Kemmer S. (Eds.) *Usage-Based Models of Language*. Stanford, CA: CSLI-Publications. 261-286.