

Corpora and the Changing Society

Studies in the evolution of English

EDITED BY

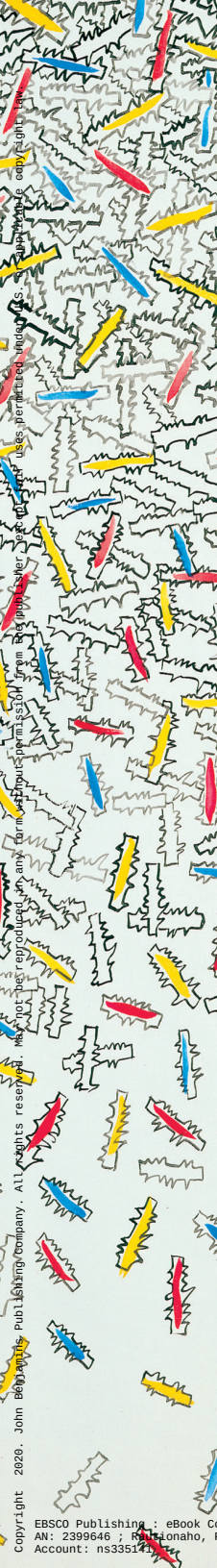
Paula Rautionaho

Arja Nurmi

Juhani Klemola

Studies in Corpus Linguistics
96

JOHN BENJAMINS PUBLISHING COMPANY



Corpora and the Changing Society

Studies in Corpus Linguistics (SCL)

ISSN 1388-0373

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

For an overview of all books published in this series, please see
benjamins.com/catalog/scl

General Editor

Ute Römer
Georgia State University

Founding Editor

Elena Tognini-Bonelli
The Tuscan Word Centre/University of Siena

Advisory Board

Laurence Anthony
Waseda University

Antti Arppe
University of Alberta

Michael Barlow
University of Auckland

Monika Bednarek
University of Sydney

Tony Berber Sardinha
Catholic University of São Paulo

Douglas Biber
Northern Arizona University

Marina Bondi
University of Modena and Reggio Emilia

Jonathan Culpeper
Lancaster University

Sylviane Granger
University of Louvain

Stefan Th. Gries
University of California, Santa Barbara

Susan Hunston
University of Birmingham

Michaela Mahlberg
University of Birmingham

Anna Mauranen
University of Helsinki

Andrea Sand
University of Trier

Benedikt Szmrecsanyi
Catholic University of Leuven

Elena Tognini-Bonelli
The Tuscan Word Centre/University of Siena

Yukio Tono
Tokyo University of Foreign Studies

Martin Warren
The Hong Kong Polytechnic University

Stefanie Wulff
University of Florida

Volume 96

Corpora and the Changing Society. Studies in the evolution of English
Edited by Paula Rautionaho, Arja Nurmi and Juhani Klemola

Corpora and the Changing Society

Studies in the evolution of English

Edited by

Paula Rautionaho

University of Eastern Finland

Arja Nurmi

Tampere University

Juhani Klemola

Tampere University

John Benjamins Publishing Company

Amsterdam / Philadelphia



The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

DOI 10.1075/scl.96

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2020002384 (PRINT) / 2020002385 (E-BOOK)

ISBN 978 90 272 0543 8 (HB)

ISBN 978 90 272 6131 1 (E-BOOK)

© 2020 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Company · <https://benjamins.com>

Table of contents

Acknowledgements	VII
Introduction: Corpora and the changing society <i>Paula Rautionaho, Arja Nurmi and Juhani Klemola</i>	IX
Part I. Changing society	
The great temptation: What diachronic corpora do and do not reveal about social change <i>Martin Hilpert</i>	3
Changes in society and language: Charting poverty <i>Gerold Schneider</i>	29
Finding evidence for a changing society: A collocational study of medical discourse in 1500–1800 <i>Maura Ratia</i>	57
Semantic neology: Challenges in matching corpus-based semantic change to real-world change <i>Antoinette Renouf</i>	79
From burden to threat: A diachronic study of language ideology and migrant representation in the British press <i>Gavin Brookes and David Wright</i>	113
Part II. Changing language	141
That's absolutely fine: An investigation of <i>absolutely</i> in the spoken BNC2014 <i>Karin Aijmer</i>	143
Two sides of the same coin? Tracking the history of the intensifiers <i>deadly</i> and <i>mortal</i> <i>Zeltia Blanco-Suárez</i>	169

So-called <i>-ingly</i> adverbs in Late Middle and Early Modern English <i>Yoko Iyeiri</i>	199
Analyzing change in the American English amplifier system in the fiction genre <i>Martin Schweinberger</i>	223
The development and pragmatic function of a non-inference marker: <i>That is not to say (that)</i> <i>Laurel J. Brinton</i>	251
Changes in transitivity and reflexive uses of <i>sit (me/myself down)</i> in Early and Late Modern English <i>Turo Vartiainen and Mikko Höglund</i>	277
Index	303

Acknowledgements

We would like to thank our co-organizers of the ICAME 39 conference, Mark Kaunisto, Minna Nevala, Päivi Pahta, Hanna Parviainen, and Paul Rickman for the smooth running of the organization process, as well as the ICAME board for their support. Needless to say, the conference was successful because of the contributions of all presenters, chairs, workshop organizers, and other ICAMers.

In the preparation of this volume, the expertise and time of the following peer reviewers was invaluable: Gisle Andersen, Alex d'Arcy, Cristiano Broccias, Robert Fuchs, Turo Hiltunen, Mark Kaunisto, Mikko Laitinen, María José López-Couso, Terttu Nevalainen, Carol Percy, Peter Petré, Heike Pichler, Robert Poole, Marco Schilk, Erik Smitterberg, Tanja Säily, Olga Timofeeva, and Rachelle Vessey. They more often than not went the extra mile in giving insightful and expert comments on the manuscripts submitted, and the resulting papers have greatly benefitted from the review process.

We also extend our gratitude to John Benjamins for including our volume in the *Studies in Corpus Linguistics* series; Ute Römer and Kees Vaes, as well as the rest of the team at Benjamins. Their support during the process has been invaluable.

Arja Nurmi's contribution for editing this volume was financially supported by the Academy of Finland funded project *Democratisation, Mediatisation and Language Practices in Britain, 1700–1950* (project number 295381).

Introduction

Corpora and the changing society

Paula Rautionahoⁱ, Arja Nurmiⁱⁱ and Juhani Klemolaⁱⁱ

ⁱUniversity of Eastern Finland / ⁱⁱTampere University

This volume presents 11 papers from the 39th ICAME¹ conference organized in Tampere in 2018, focussing on the theme of the conference, corpus linguistics and the changing society. It is a truism that both language and society are constantly changing. It is also commonly accepted that changes in society are reflected as changes in language and vice versa. When new technologies are introduced, we gain new vocabulary to discuss them. When our methods of communication change, so do our linguistic means of communicating; when the ideas and phenomena change and evolve, so does the language used to describe them. Words change their meaning as the entity they denote changes. Increasing gender equality, for instance, has led to changes in vocabulary referring to men and women (e.g. *chairman* vs. *chairperson*). Grammatical change may be more subtle, but nonetheless able to reflect cultural change; the decline in frequency of the English core modals and the concomitant increase of semimodals has been linked to the process of democratization. While we sometimes think of social and societal change as phenomena of the present day and language change as something only observed in the past, both are and have been present for as long as there have been languages and societies.

This volume approaches both patterns of change, those taking place on the level of language and those where social and linguistic change are reflected in each other. All papers included deal with English, choosing anything between a long diachrony from Old English to the present, or a timespan of a few decades in the immediate past. Two unifying threads can be seen again and again in the studies presented here: the ever-increasing availability of large diachronic corpora and the growing sophistication of corpus-linguistic methodology. In true ICAME tradition, the authors of the papers are aware of the potential pitfalls of the data

1. ICAME is the International Computer Archive of Modern and Medieval English, an international organization of linguists and information scientists working with English machine-readable texts. See <<http://clu.uni.no/icame/>> for more information.

and methods they use, and evaluate both as part of their linguistic inquiry into the phenomena they are studying.

The studies included in this volume approach the connections between language change and societal change with the tools offered by corpus linguistics. While the approach is by no means a new one, the advances in corpus methods and the availability of new datasets have made it possible to ask new questions and expand our understanding of different aspects of language and language change. The increasingly sophisticated statistical methods used to explore modern, extensive corpora allow for insights into questions previously studied with the help of small corpora, while at the same time entirely new questions can be explored with the help of these tools. The cross-fertilization of cultural studies and corpus linguistics, we hope, is beneficial for both parties, corpus linguists and researchers of cultural change alike. Corpus linguistics offers a vast array of materials and methods to investigate cultural and societal change, while cultural studies provide the theoretical background on which to build our research. The studies included in the present volume showcase the potential avenues of combining changing language and changing societies.

The main body of the volume is divided into two parts, reflecting the two main approaches taken by the authors; some focus on the changing society, while others focus on the changing language. Part I, *Changing society*, consists of five papers which explicitly discuss the connections between changes in society and the ways in which this is reflected in language change and language use. The papers also raise questions regarding the reliability of corpus-based data and corpus methodology in identifying such connections.

Martin Hilpert starts by asking a question that is very relevant to the present volume: do “frequencies of linguistic items reflect their occurrence in the real world?” – is there a correlation between how we use language and the real world? Detecting new correlations between linguistic and social phenomena is ever more tempting with the abundance of resources today, but Hilpert draws together some of the pitfalls in the analysis of (diachronic) corpus data. He then goes on to suggest measures to be taken in order to avoid, for instance, reporting spurious correlations, and reminds the reader of the importance of theoretical considerations in setting up a study.

Gavin Brookes and **David Wright** examine right-leaning press representations of non-native English-speaking migrants living in the UK. They identify where and how these representations have changed, or remained stable, over the period 2005 to 2017. The “speak English” debate emerges as a dynamic issue; during the period of time under investigation, the attitudes towards non-native English-speaking people shifted from an economic burden to a threat to the success of the local native groups. Brookes and Wright argue that the way that the right-leaning media

represents this group of people ultimately affects the public opinion and government policy regarding migrants in the UK.

Antoinette Renouf combines a large corpus of newspaper data with a test of corpus-linguistic tools for tracking lexical innovation through collocation. She highlights the potential advantages as well as the disadvantages of the corpus tools AVIATOR, ACRONYM, and APRIL as means of identifying semantic change in the meaning of words, particularly in the culturally, socially, and politically driven changes in recent decades. She tests five different words (*birther*, *normalis/zation*, *cougar*, *snowflake*, and *ghosting*) which have gained new meanings and discusses the ways of identifying these meanings and the pitfalls of semi-automatic identification.

Gerold Schneider uses the topics of poverty and industrial revolution as a case study to show how societal and linguistic changes can be detected using historical corpora. Building up a repository of statistical approaches, for instance topic modelling, Schneider shows how poverty, industrial revolution, and urbanization are associated with one another through, for instance, the associations of war, religion, and family.

Taking a collocational approach, **Maura Ratia** explores the changing attitudes of medical practitioners towards their patients in the years 1500 to 1800. By investigating the collocates of the words *patient* and *patients*, Ratia is able to detect a growing concern of the patients' well-being and suffering in the 18th century.

Part II of the volume, *Changing language*, consists of six papers focusing on individual linguistic features and their development during a specific historical period, ranging from Old English to the 21st century. There is a particular focus on various types of adverbials, but changes and variation in the verb phrase are also discussed.

Karin Aijmer investigates the meanings and functions of the intensifier *absolutely* in recent British English. Her results show that the developments undergone by *absolutely* can be described as a trajectory from degree modifier to emphazier and to discourse marker. A sociolinguistic account of the variables age and gender indicate that, perhaps contrary to expectations, it is the older speakers and male speakers who use *absolutely* more frequently than younger or female speakers.

Focusing on two death-related intensifiers, *deadly* and *mortal*, **Zeltia Blanco-Suárez** tracks their history from Old English to Present-day English. A collocational analysis reveals that these intensifiers have evolved from their original literal meaning ('cause death') to the semantically bleached meaning ('very'). Blanco-Suárez concludes that *deadly* and *mortal* may well be victims of fashion: since Late Modern English they significantly decreased in frequency, and were replaced by other more powerful intensifiers.

Yoko Iyeiri focuses on adverbs ending in *-ingly*, tracking their history from Late Middle to Early Modern English, and finds that not only their frequency and

types, but also functions have expanded. Iyeiri argues that the lexicalization of the *-ingly* adverbs may have been propelled by the creativity of individual writers in adopting and forming new adverbs. Finally, the so-called “Harry Potter adverbs” seem to have been in use earlier than what has been attested in existing research.

Martin Schweinberger’s study investigates the diachronic development of amplification of adjectives in American fiction data. His results indicate that *so* has replaced *very* as the dominant amplifier in predicative contexts in the 1990s. However, despite their general semantic similarity, *so* and *very* still exhibit distinct semantic profiles with respect to their semantic specialization. Schweinberger thus argues that it would be inaccurate to speak of full lexical replacement.

The topic of **Laurel J. Brinton**’s study is the development and functions of *that is not to say (that)*, an introductory clause refuting an inference that could be drawn from the previous discourse. Combining diachronic and synchronic approaches, Brinton tracks the frequency and pragmatic functions of this non-inference marker from the 16th century to the present day. The developments could be seen as the result of grammaticalization, but Brinton points out that some trends are actually more indicative of colloquialization.

Bringing the volume to its conclusion, **Turo Vartiainen** and **Mikko Höglund** focus on the historical development of a single verb, *sit*, taking stock of the variation between the simple and the SELF reflexive strategies in the Early Modern period, and the transitivization of *sit down* in Late Modern English. The study is one of the first to apply the Transitivity Hypothesis to historical data, and shows the usefulness of the hypothesis in explaining paths of transitivization in the evolution of English.

As the range of studies included in this volume illustrates, there are multiple ways in which to approach language and the changing society. The new insights gained through the application of up-to-date methodologies to modern, extensive datasets showcase the potential of this combination of viewpoints. Whether focusing on the interconnectedness of linguistic and social change or on the level of linguistics alone, the studies in this volume hopefully advance our understanding of not only the individual topics, but also more broadly the advantages of combining different outlooks and perspectives in new and innovative ways.

PART I

Changing society

The great temptation

What diachronic corpora do and do not reveal about social change

Martin Hilpert

University of Neuchâtel

This paper examines the potential large diachronic corpora hold for the study of social change. Resources such as COHA or *Google Books* allow us to detect shifts in the frequencies of linguistic elements, which can then be interpreted as reflections of developments in society. This paper addresses the practicalities of this question in two parts. The 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 the English *make*-causative over the past 150 years. Examining the variables of animacy and verb semantics, 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 *Corpus of Historical American English* (COHA; Davies 2010) or *Google Books* (Michel et al. 2011) 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. (2011: 176), interpreting frequencies from diachronic corpora as reflections of on-going 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 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 bring the evidence to light. It is this very tempting idea that the title of this chapter 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 have been used to explore social change. Greenfield (2013) takes positive frequency trends of word forms 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 frequency trends of linguistic items 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 Books* corpora for these items. 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.

It is readily apparent that the three examples of *Gemeinschaft*-adapted lexis are on the decline, while the reverse is true for *Gesellschaft*-adapted words, which is 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

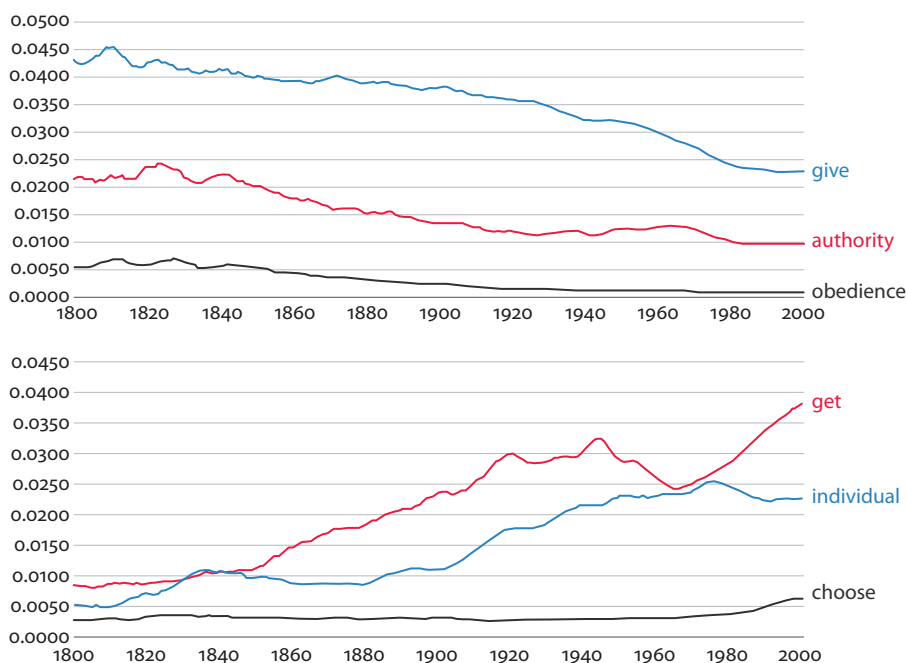


Figure 1. Relative frequency trends for six lexical elements in the *Google Books* corpora

the corpus-linguistic research community would be hesitant to generalize from a corpus, large as 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 chapter: 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 chapter 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 and construction-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.

It goes without saying that this list is not exhaustive. Differences in corpus size, changes in textual genres, and the uneven representation of different varieties and registers present problems that also need to be taken into account. Still, the discussion in this paper will focus on the five points that are listed above. The second part of this paper implements these ideas in a study of change in the English *make*-causative construction (Gilquin 2010; Kemmer 2001). 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 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 lexical 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 some predictions that can be derived from the hypothesis of social change. At the same time, the results deviate in several respects from Greenfield's empirical observations, so that a common underlying cause for both developments is ultimately unlikely.

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:

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 *News on the Web* corpus (NOW; 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 real-world use of screwdrivers, which, of course, 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.

Table 1. Verbal elements co-occurring with *with a screwdriver*

Collocate	Frequency	Collocate	Frequency
<i>was</i>	105	<i>been</i>	14
<i>armed</i>	80	<i>being</i>	13
<i>stabbed</i>	70	<i>forced</i>	13
<i>threatened</i>	42	<i>open</i>	13
<i>attacked</i>	33	<i>stab</i>	11
<i>said</i>	30	<i>found</i>	10
<i>were</i>	18	<i>can</i>	10
<i>be</i>	17	<i>is</i>	10

A reasonable objection to this counterargument would be that the NOW corpus contains writing from news texts, and that ordinary events that take place within the private sphere, 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 in news media despite being fully expectable, and while the specific outcome of an event such as a football tournament may be unpredictable, many other aspects, including the participation of important players and the venues of particular matches, are not.

The NOW corpus further allows comparisons between events that are equally undeserving of media coverage, and that hence should be used with frequencies that match their frequency of occurrence in the real world. This expectation is not always met. For example, the strings *doing my laundry* and *doing my taxes* occur with virtually identical frequencies in the corpus, 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 and construction-specific frequencies

In a response to Greenfield (2013), Flach (2013) points to the large body of work in corpus linguistics that indicates 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 *get* 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 (2013) 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 hand-picked lexical material in corpus-based studies of social change. Some of the resulting problems are attenuated if grammatical constructions are used instead for that purpose. One 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 as a more reliable source of data for the analysis of social change.

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? Unfortunately, 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 that has no explanation in terms of any factual 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. 2011), 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 positive 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 positive correlation between lexical diversity in the investigated languages and global mean sea level, a variable that has no influence whatsoever on language use. This means that there is a problem. How can spurious correlations be distinguished from true correlations? At a very basic level, the study of correlations should be theoretically informed, so that the analysis is grounded in an independent, external motivation. Yet, this leaves open the possibility that a plausible correlation is actually spurious.

The safeguard that Koplenig and Müller-Spitzer (2016: 6) offer against falling into the trap of spurious correlations in diachronic data is the practice of correlating

the values of changes between temporally adjacent periods, rather than correlating the absolute frequency values of those periods. The graphs in Figure 2 illustrate why this is important. The three panels in the upper row show the frequency developments of three fictive linguistic phenomena A, B, and C. The frequencies for these phenomena have been created artificially. The x-axis shows the passage of time; the y-axis shows values of increasing text frequency. A standard correlation test, such as Pearson's r or Kendall's tau (Hilpert & Gries 2009) will confirm the visual impression that the frequency developments are strongly correlated. Koplenig and Müller-Spitzer caution against the use of correlation statistics on the basis of frequency values. The reason for their argument becomes apparent in the three panels in the middle row, which show the changes in frequency that each phenomenon

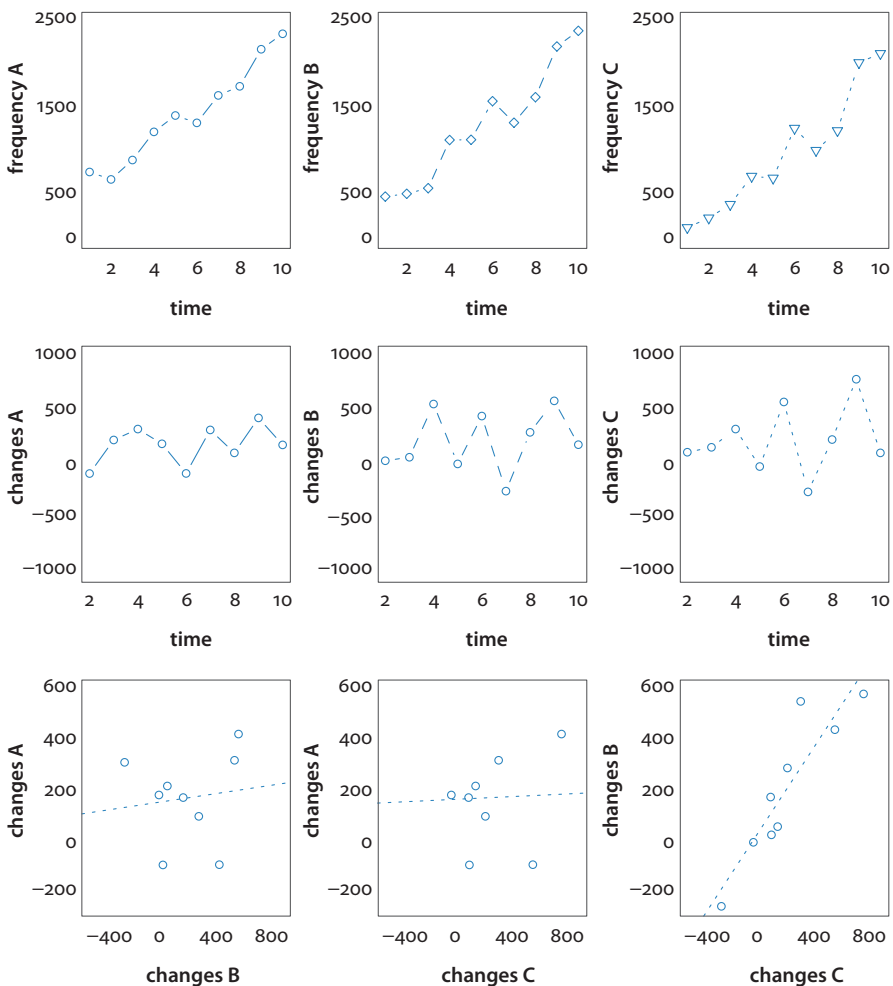


Figure 2. Frequency developments in three phenomena A, B, and C

undergoes from one temporal period to the next. Points above zero mean that the phenomenon has increased in frequency between two periods. Drops do not necessarily mean that the phenomenon has decreased in frequency. As long as the drop does not fall below zero, the phenomenon has still gained in frequency, but just not as much as before. A visual comparison of the three panels reveals that phenomena B and C are rather similar with regard to their changes, whereas phenomenon A exhibits changes that appear to be unrelated to the changes in B and C. This impression is confirmed by the graphs in the bottom row of Figure 2. The first two panels show that changes in phenomenon A do not correlate with changes in either B or C. The third panel, by contrast, shows that any change in B for a given period strongly predicts the kind of change that will be observed in C during that time. The changes are not completely identical, but they line up rather closely in terms of direction and magnitude.

Koplenig and Müller-Spitzer (2016: 6) apply this reasoning to the apparent correlations of lexical diversity with population size and sea levels, and they show that if year-to-year changes are taken as the relevant measure, there is no longer a statistically significant correlation. The idea of accepting correlations only when changes correlate acts as a powerful filter against potentially spurious correlations that may appear in large databases (Roberts & Winters 2013). Another look at Figure 1 and the decreasing frequency trends of *give*, *authority*, and *obedience* actually shows that some peaks and troughs line up more or less, which would support the idea that their decline is motivated by a common cause. For the increases of *get*, *individual*, and *choose*, on the other hand, it is not possible to make that argument. Their frequency increases show very different profiles and are thus unlikely to survive a correlation test that takes year-by-year changes as the basis for a comparison. Nor would there be any reason to conclude that the decreases of *give*, *authority*, and *obedience* are in any way related to the increases of *get*, *individual*, and *choose*. In summary then, the technique suggested by Koplenig and Müller-Spitzer casts doubt on the argument put forward by Greenfield (2013).

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

Up to this point, changing frequencies in historical corpus data have been discussed as a reflection of changes in society. Importantly, changes in diachronic corpora may be due to several other reasons, including of course linguistic change and specifically phenomena of grammaticalization (Hopper & Traugott 2003). Mistaking one type of change for the other can lead to conclusions that are unwarranted. For example, the rise of the *be going to* future construction does certainly not mean that speakers of English moved around to a greater degree. But how can it be decided whether a given development in the linguistic record is evidence for social change or for

linguistic change? Szmrecsanyi (2016) addresses this issue in a diachronic study of the English genitive alternation during Late Modern English. Szmrecsanyi cautions against taking changes in text frequency as a measure and instead proposes that grammatical change is best observed in terms of shifts in the ecology of probabilistic choices between alternative constructions (2016: 154). To give a concrete example, grammatical change would be observed if the factors that lead speakers to choose either the English *of*-genitive or the *s*-genitive are different across two different periods of historical corpus data. Using regression modelling, Szmrecsanyi (2016) shows that with regard to the English genitive alternation, some conditioning factors indeed undergo diachronic change. Whereas an inanimate locative possessor (as in *the inhabitants of the island*) very strongly biases speakers towards the *of*-genitive in the early 19th century, that bias has become a lot weaker by the end of the 20th century, so that it is relatively more common for a present-day speaker to produce a phrase such as *the island's inhabitants* (Szmrecsanyi 2016: 168). Since the statistical assessment of the conditioning factors controls for the text frequencies of the participating constructions and the values of their formal and functional variables, the analysis remains robust, even when other aspects of the corpus data are in flux.

Szmrecsanyi's argument mainly addresses a potential pitfall in the analysis of grammatical change, but his general conclusion holds for the study of social change as well. Changes in text frequency can be brought about by a wide variety of factors, which renders their interpretation rather difficult. A more promising instrument for the study of change, be it social or linguistic, is the observation of speakers' choices, given a set of alternatives. This insight completes our theoretical discussion of problems that studies of social change on the basis of diachronic corpus data have to reckon with. While it is clear that the obstacles for a successful undertaking are substantial, this should not keep us from asking research questions that try to get at social phenomena through the analysis of linguistic data. The next section will discuss how the English *make*-causative construction may be analyzed for the purpose of studying social change in a way that meets the theoretical challenges that were outlined above.

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

The starting point for the empirical part of this paper is Greenfield's (2013) observation that words such as *authority* and *obedience* are in decline, coupled with the relatively uncontroversial claim that the notions that are expressed by these words have become less important in English-speaking societies. Thankfully, environments such as the workplace, the family, academia, or schools have become much less hierarchical. It is a reasonable hypothesis that this social development

should in some ways be reflected in shifting patterns of language use. The analysis in this section will investigate whether the diminishing social value of interpersonal authority can be observed in the usage of a grammatical construction, namely the English *make*-causative (Gilquin 2010; Kemmer 2001).

The choice of this particular case study is motivated by the theoretical challenges that have been outlined in the previous section. The analysis takes to heart Szmrecsanyi's (2016) advice that changes in text frequency may not reveal the whole picture. Instead of looking at changes in the text frequency of the *make*-causative, the focus of the analysis will be on changes in the relative frequencies of different semantic variants of the construction. This approach will also address the problem of potentially non-equivalent semasiological and onomasiological frequencies. While it is certainly useful to determine how often speakers of English use the *make*-causative, the main aim of the present analysis is to determine for what purpose speakers of English use the construction when they are using it. Another important part of the study will be an examination of the verbal collocates of the *make*-causative. The construction holds a slot for a lexical verb in the infinitive (*It made me think*), and changes in the construction's collocational preferences can reveal how the construction has changed semantically over time (cf. Hilpert 2006, 2008). By looking at the lexical collocates that occur in the context of a specific construction, the problem of a potential subjective bias in selecting lexical items for analysis is avoided. The basis for the analysis are simply all verbs that occur in the corpus with the *make*-causative construction. And it is not the overall text frequency of those verbs that informs the analysis, but rather the frequency with which those verbs occur within the construction.

A final reason for choosing this construction as an object of analysis is that a parallel case from Dutch provides some food for thought. Verhagen (2000: 263) compares two Dutch constructions with the causative verbs *doen* and *laten*, which are structurally very similar to the English *make*-causative, and which are exemplified below.

- (2) a. De sergeant liet ons door de modder kruipen.
the sergeant let us through the mud creep
'The sergeant made us crawl through the mud.'
- b. De stralende zon doet de temperatuur oplopen.
the shining sun does the temperature rise
'The bright sun makes the temperature rise.'

Verhagen retrieves examples of these constructions from diachronic corpora and finds that for *doen*, animate causers are on the decline. He explains this in the following way (2000: 272):

Now, of the factors mentioned above, the most plausible one to have changed much over the last three centuries is that of authority: We already know from all kinds of sources that 200 years ago, authority was a much more important determinant of social and personal relationships, or at least of their evaluation, than it is today.

Drawing on Verhagen's empirical findings for Dutch, it can be hypothesized with regard to English that fading authority has left its mark on the way the English *make*-causative has been used over the past two centuries. Three specific predictions can be derived from that general hypothesis. The first prediction would be that uses of the *make*-causative that express situations of interpersonal authority should decrease. A second prediction is that uses of the *make*-causative with animate causers should decrease, since these uses represent at least partly those that are affected by the first prediction. Lastly, it can be predicted that uses of the *make*-causative in which both causer and causee are animate should decrease, since the meaning of interpersonal authority requires both parties to be animate. With these predictions in place, it is time to move on to a more detailed description of the *make*-causative construction.

3.1 The English *make*-causative construction

The English *make*-causative is a sentence-level construction that involves a subject, a form of the verb *make*, and an object of *make* that simultaneously functions as the subject of a following verb in the infinitive (Gilquin 2010: 65). In the following, the subject will be referred to as the causer, while the object of *make* will be called the causee. The construction is illustrated by the following examples, appearing in the lyrics of popular songs.

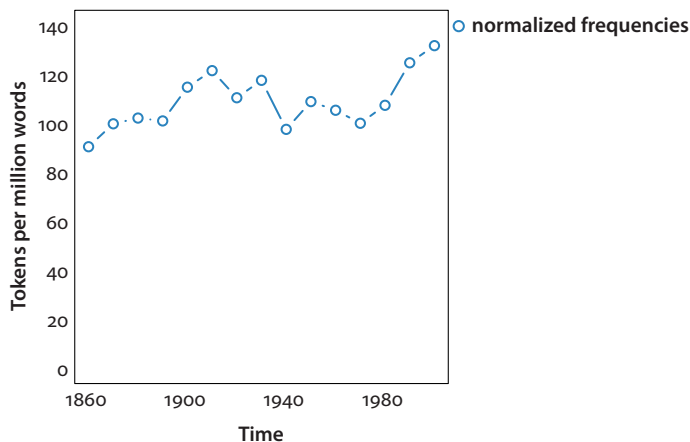
- (3) a. You *make* a grown man cry.
- b. I sing the words that *make* you blush.
- c. You *make* the rockin' world go round.
- d. It *made* my eyes be gettin' blurry.

The examples show that both causer and the causee may vary with regard to animacy. Both are animate in (3a); Example (3c) shows an inanimate causee; in Example (3d) the causer is inanimate. What can be said about other semantic aspects of the *make*-causative? Kemmer (2001) identifies three semantic frames that are conventionally associated with the construction. These frames are respectively called the mechanical action frame (*The sun made the ice melt*), the experiential response frame (*It made her feel good*), and the compulsion frame (*Don't make me marry him*). It is the latter that represents the verbalization of interpersonal authority. The *make*-causative construction can be regarded as a grammaticalized means

for the expression of compulsion. Kemmer points out that the compulsion frame is characterized by a lower type and token frequency of collocating lexical verbs, but that those verbs are semantically more diverse than the verbs that are found in both the mechanical action frame and the experiential response frame. Gilquin (2010: 183) further observes that the construction evokes two semantic fields that are partially overlapping, namely field of the physical self and the field of the mental self. A verb such as *itch* describes a bodily reaction and thus relates to the former, whereas a verb such as *think* describes a cognitive process and relates to the latter. Among the particularly frequent collocates of the *make*-causative construction, Gilquin (2010: 184) notes the verbs *feel*, *laugh*, *look*, *think*, *know*, *wonder*, *appear*, and *seem*. The empirical investigation in this paper builds on these insights. The next section discusses the data that were retrieved for the analysis.

3.2 Corpus data and descriptive statistics

For the analysis, examples of the *make*-causative were taken from COHA (Davies 2010), using the data representing the 15 decades from the 1860s onwards. Examples were retrieved searching for a form of *make*, followed by a noun phrase with up to three words and a verb in the infinitive. After manual clean-up, a concordance of 38,597 examples remained. In a first analytical step, those examples were annotated for the lexical verb in the infinitive. The dataset holds 1,681 different verb types. The examples were further manually annotated for the animacy of causer and causee. A simple binary coding of animacy into animate and inanimate was adopted (cf. Zaenen et al. 2004 for a more detailed approach). Problematic decisions include human collectives, which were coded as animate, or body parts, which were coded as inanimate. Figure 3 offers an overview of different frequency developments of the *make*-causative in COHA.



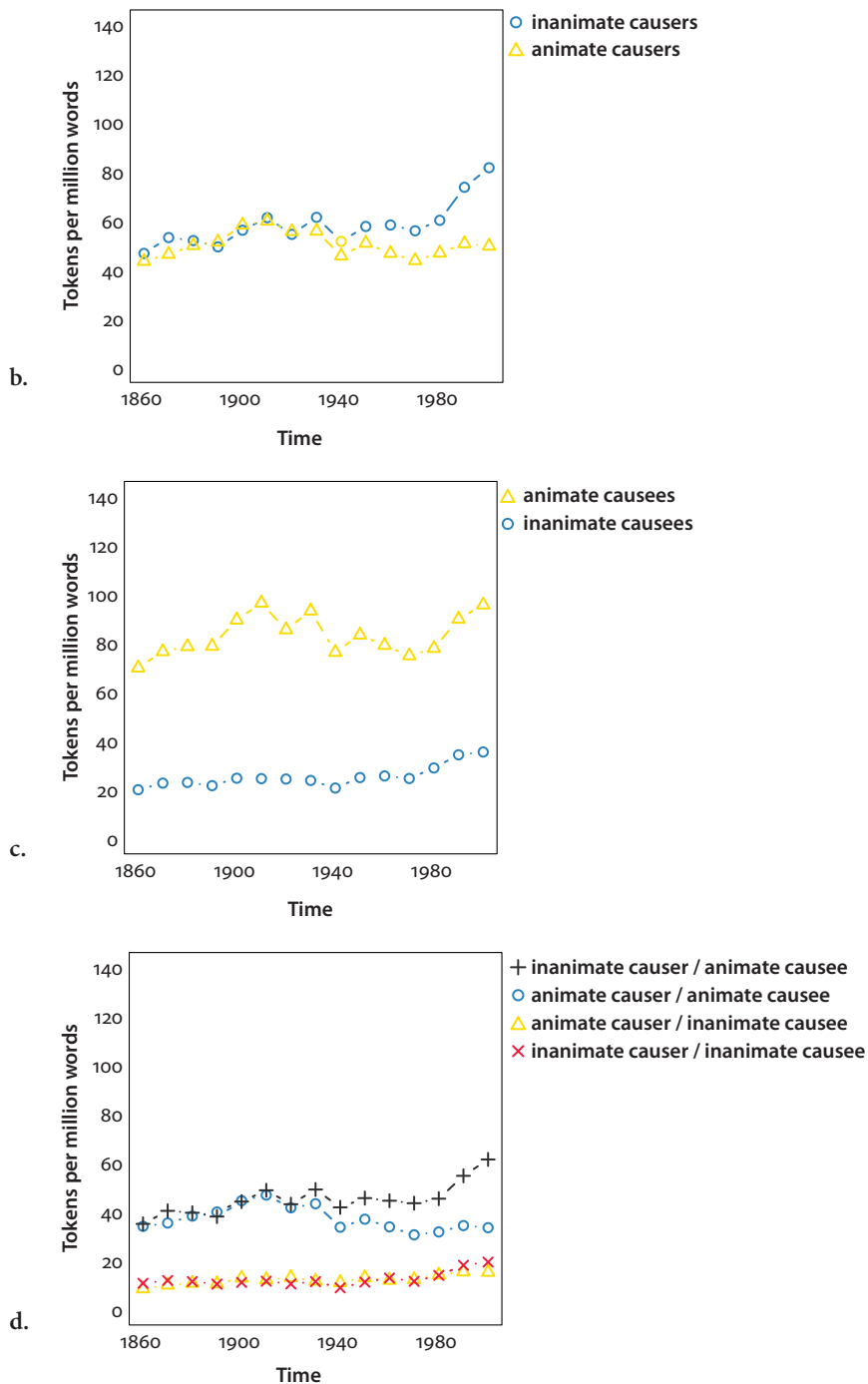


Figure 3. Frequency developments of the *make*-causative in COHA (1860–2000)

A first result that can be seen in the first panel is that the *make*-causative actually undergoes a moderate frequency increase in COHA. This is in partial contrast to the Dutch *doen* causative, which decreases in text frequency over time (Verhagen 2000: 274). The second panel shows that the observation that can be made is that animate causers are more or less stable, while inanimate causers are on the rise. This means that the second prediction that was made earlier is true only in relative terms. The relative frequency of animate causers decreases, but their absolute frequency actually stays the same. The third panel shows the frequency developments of animate and inanimate causees. Here it appears that both animate and inanimate causees are on the rise. What is crucial of course is the development of the four different configurations of causer and causee animacy. The frequencies of these configurations are shown in the fourth panel. This graph shows a frequency increase of the configuration of inanimate causers and animate causees. This configuration is present in examples such as *The nap made him feel refreshed*. It is this type of usage that has been diachronically on the rise. The lower part of the graph shows the almost parallel developments of the two configurations with inanimate causees. Both of these remain relatively infrequent, but show moderate increases over time. The only configuration that does not increase and that even shows a downward trend over the past century is the configuration of animate causers with animate causees. This is in line with the third prediction that was made above. All-animate uses of the *make*-causative include those uses that express relations of authority and obedience, and these uses are historically on the decline. However, all-animate uses can of course instantiate the experiential response frame (*You make a grown man cry*) as well as the compulsion frame (*Don't make me marry him*). In order to address the first prediction and to find out what is really happening semantically with the *make*-causative, it is necessary to take a closer look at the verbs that are used in the construction. The following section will describe how a distributional approach can be put into the service of studying the semantic development of the *make*-causative construction.

3.3 Using distributional semantics to study the development of the *make*-causative

Distributional semantics (Turney & Pantel 2010) is an approach to word meaning that measures semantic similarity between different linguistic items in terms of the collocational contexts in which those items are typically found. Words with related meanings are expected to have similar sets of collocates. For example, in a large corpus of English such as the NOW corpus (Davies 2013), the noun *guitar* will frequently occur in the context of words such as *playing*, *acoustic*, *music*, *sound*, or

piano. Semantically related nouns will share some of these collocates. Searches for *drums* or *bass* reveal that these nouns have similar collocational profiles. By contrast, a semantically unrelated word such as *lizard* will occur with a very different set of collocates. By measuring mutual semantic similarities between pairs of lexical items that form a larger set of words, it is possible to map out a semantic space that allows the detection of semantic clusters and categories within that set of words. The analysis in this section uses data from the *Corpus of Contemporary American English* (COCA; Davies 2008) in order to compare the meanings of the lexical verbs that are found in the *make*-causative construction. As was discussed in the previous sections, some of these verbs denote experiential responses (*feel*, *cry*, *blush*), others refer to mechanical actions (*melt*, *fall*, *crumble*), and still others verbalize a wide range of coerced activities (*marry*, *work*, *wash*). The goal of the analysis is to find out how the semantic space of the verbs in the *make*-causative construction has changed over time, and whether the declining importance of interpersonal authority can be observed in those changes. Studies in which distributional semantic methods have been applied to questions of language change include Sagi, Kaufmann and Clark (2011), Hilpert and Perek (2015), Hilpert (2016), and Perek (2016). The following paragraph briefly describes the methodology that has been adopted for the present analysis. A more detailed account of the analytical steps can be found in Hilpert and Correia Saavedra (2017).

For the analysis of the verbs in the *make*-causative, a random subset of 29 million words from the offline version of COCA (Davies 2008) was used. Concordances with a context window of four words to the left and four words to the right were retrieved for each of the 1,681 verb types in the *make*-causative. These concordances were cleaned from stop words and transformed into frequency vectors that represent the collocational profiles of the lexical verbs. The frequency vectors of all verbs were integrated into a collocation matrix. Context items that did not occur more than 10 times in the entire matrix were excluded. On the basis of the collocate frequencies between each verb and each collocating context item, values of collocational strength were determined with the measure of pointwise mutual information (Turney & Pantel 2010: 157). Context items that did not show a pointwise mutual information value of at least 5.5 for at least one verb type were excluded. The end result of this process was a collocation matrix with 1,681 verb types in the columns and a total of 4,970 different context items in the rows. For the purpose of visualization, the collocation matrix was transformed with the cosine, and metric multidimensional scaling was used to create a two-dimensional map of the verbs that are used in the *make*-causative. Figure 4 visualizes the results of that procedure.

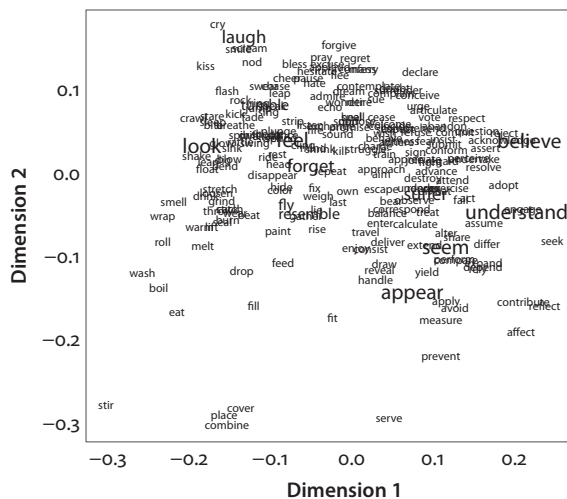


Figure 4. A semantic vector space of the verbs in the *make-causative* (1860s)

In the graph, verbs that appear in the *make-causative* construction are arranged according to their collocational profiles in COCA. Verbs with similar sets of collocates are placed in close mutual proximity; verbs with very different collocates are placed further apart. The font size of the verbs reflects their frequency of usage in the 1860s data of COHA. In that period, the most frequent verbs are *look* and *feel*, followed by *believe*, *seem*, *appear*, and *understand*, which is consistent with the results obtained by Gilquin (2010: 184) on the basis of more recent data. Of special interest for the present discussion is not just the frequency of these verbs, but rather their distribution across the graph. The upper left of the graph shows a cluster with the verbs *cry*, *laugh*, and *smile*, which clearly instantiate the experiential response frame. Towards the lower right, the graph shows several verbs that are abstract, like *appear*, *seem*, *reflect*, and *affect*, and a bit further up there are cognitive verbs such as *understand* and *believe*. All of these have to do with epistemicity, that is, states of knowledge. The graph thus nicely corroborates Gilquin's (2010: 183) argument that the verbs in the *make-causative* cover a continuum between verbs that relate to the physical self and verbs that relate to the mental self. The verbs that are higher up emphasize the human cognizer, whereas those further below verbalize epistemic relations in which the human cognizer is only implied. The graph further shows a number of verbs that relate to the compulsion frame. This includes verbs such as *wash* or *eat* in the lower left corner but also *paint* or *feed*, which are located towards the center, as well as *confess*, which is to be found towards the upper area of the graph. Naturally, the verbs encoding actions that can be brought about through interpersonal authority are semantically of a more varied nature than emotional or cognitive responses, and are therefore not arranged in a neat cluster.

On the basis of distributional data, it is possible to determine the areas in a semantic space that show particularly strong developments over time. With regard to the *make*-causative, it can be investigated in what semantic areas many new verbs appear or existing verbs become more or less frequent. In Figure 5, the data from COHA is shown in four sequential periods. Each panel shows the verb types that are attested during a given time period. As in Figure 4 above, the positioning of the verbs is based on the collocate profiles from COCA, so that each verb remains constant in its place from one panel to the next. What changes is that new verb types may appear, or old types may disappear. Also, the density lines in the background of the panels give an indication as to how the areas of the semantic space differ with regard to the token frequency of the *make*-causative. Peaks are due either to single verbs that are particularly frequent (e.g. the verb *believe* in the second period), or they represent semantic areas that are densely populated by a large number of types (e.g. the area around the verb *understand* in the second period).

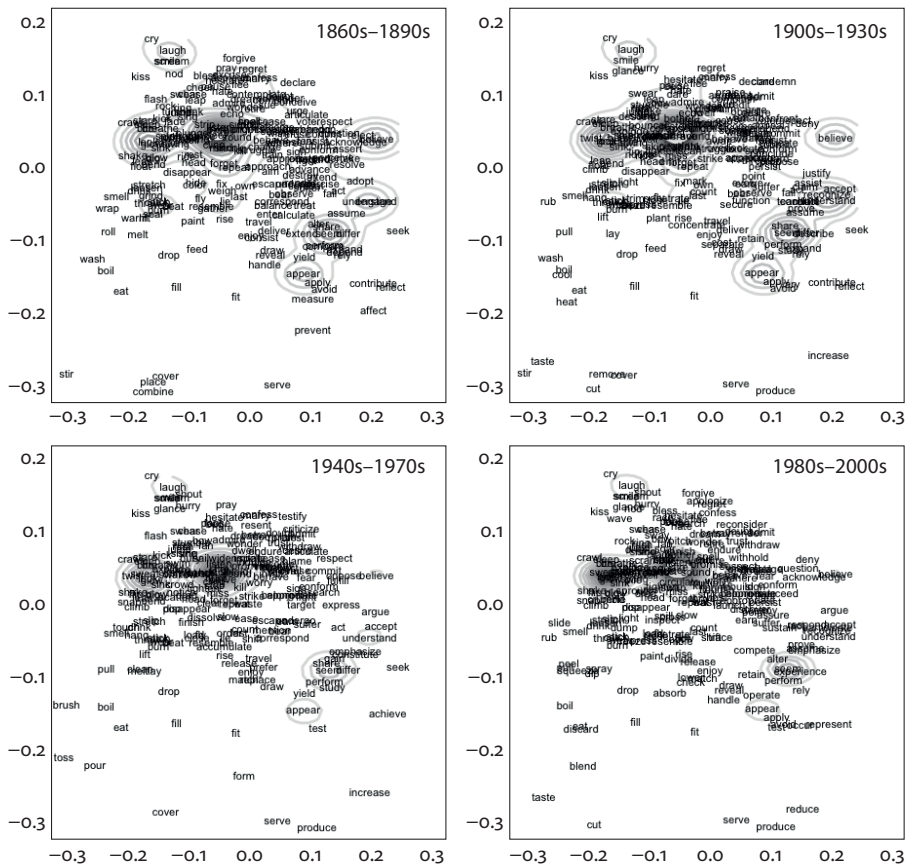


Figure 5. Changes in the semantic vector space of the *make*-causative (COHA 1860–2000)

The graphs in Figure 5 allow three observations. The first is that the overall collocational profile of the *make*-causative has remained relatively stable over the past 150 years in terms of its verb types and their token frequencies. All four panels show a densely populated area of verbs that stretches from the upper left half of the graphs to the lower right. That area represents a semantic continuum from physical experiential responses (*cry*, *laugh*) to emotions (*feel*, *admire*), to cognitive processes (*believe*, *accept*), and further to epistemic predicates (*appear*, *reflect*). These verb types embody the most important semantic aspects of the *make*-causative. The second observation is that some areas remain constantly high in token frequency, whereas others see decreases. Two areas that are particularly high in token frequency are around the verbs *look* and *feel* (cf. Figure 4). By contrast, the verbs *believe*, *understand*, and *appear*, which are highly frequent in the first two periods, decrease in frequency over time. Thirdly, the graphs show some fluctuation of infrequent verb types that are only present in one or two periods (*heat*, *discard*, *reduce*, *testify*, and others). The graphs do not support the idea that the *make*-causative construction has become any less frequent as an expression of interpersonal authority. Verbs that fit the compulsion frame, such as *confess*, *accept*, or *serve*, are still attested in the fourth period, and the wide semantic distribution of such verbs makes it difficult for speakers to recognize any conventional limitation on how the construction can be used. For the main hypothesis of this study, this means that the fading of authority has impacted the use of the construction only in minor ways.

In order to be able to assess this point more thoroughly, it is useful to present the same data from a slightly different perspective. Figure 6 shows again the four periods and their respective verb types, but instead of visualizing token frequencies through density lines, the font size represents token frequencies of the verbs and the font color represents their relative preference for the configuration of an animate causer and an animate causee. The darker the font of a verb, the more it is used with the all-animate configuration that lends itself to the expression of interpersonal authority and compulsion. Verbs shown in black occur at least 75% of the time with an all-animate configuration. Verbs in the lightest grey occur at most 25% with the all-animate configuration. In the first period, the verbs *believe* and *understand* show a strong preference for all-animate uses. Between the second and the fourth period, these verbs decrease in frequency and also in their preference for the all-animate configuration. Crucially, *believe* and *understand* do not instantiate the compulsion frame. While causees can be actively forced to *confess* a crime or to *eat* their vegetables, acts of believing or understanding are based on the independent agency of the causee.

In summary, the graphs in Figure 6 suggest that the overall frequency decrease of the all-animate configuration (cf. Figure 3, panel 4) is not due to the demise of the compulsion frame. Rather, the compulsion frame was already marginal in the *make*-causative in the 19th century, and it has remained marginal since then.

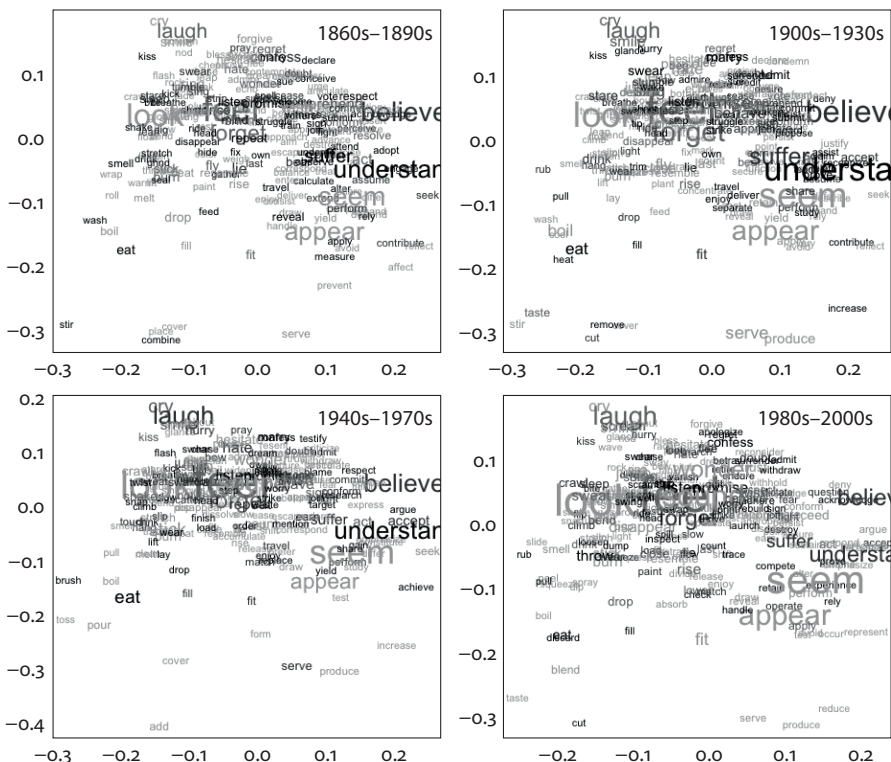


Figure 6. Relative preference for animate causers and animate causees

3.4 Discussion

The present analysis was informed by the hypothesis that social change, specifically the fading of interpersonal authority, influenced the way the English *make*-causative has been used over the past two centuries. The hypothesis gave rise to three predictions, which can now be assessed. Frequency counts from COHA confirm two of those predictions, namely that uses of the *make*-causative with animate causers are on the decline, and that the construction is used less and less with the configuration of an animate causer and an animate causee. Both of these observations are in line with the general hypothesis. At the same time, the analysis of diachronic changes in a semantic vector space that represents the verbs occurring in the *make*-causative construction fails to support the prediction that the construction declines in its use of expressing compulsion. Rather, it has been shown that the frequency decrease in all-animate uses of the construction is mainly due to verbs that do not instantiate the compulsion frame.

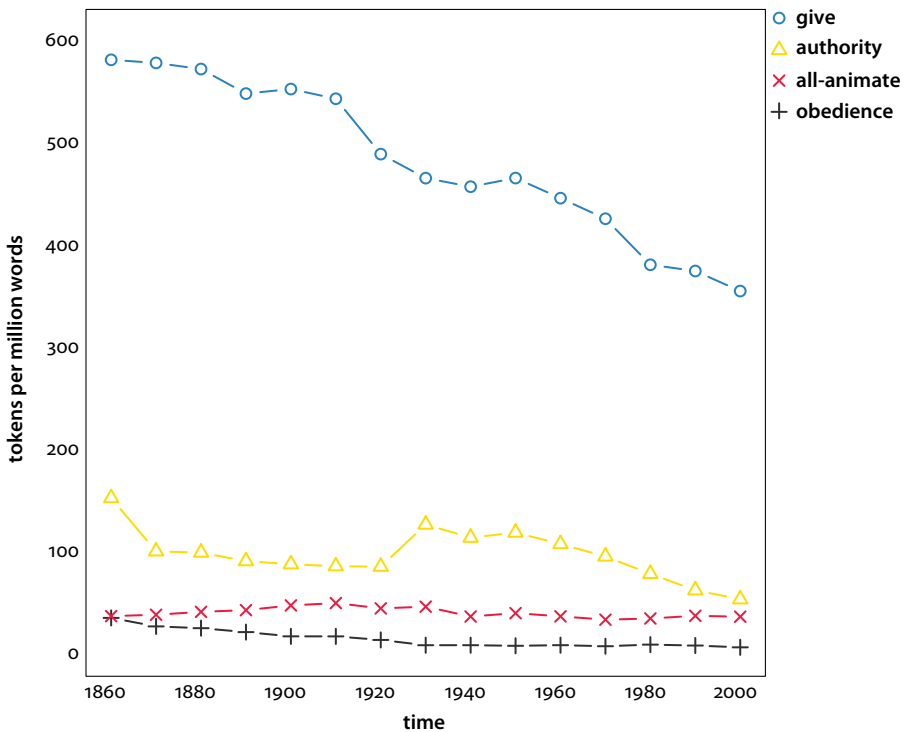


Figure 7. Frequency decreases in COHA: *give*, *authority*, *obedience*, and all-animate uses of the *make*-causative

Another piece of evidence can be gleaned from Figure 7. The graph shows data from COHA that corroborate Greenfield’s (2013) observation that the text frequencies of the words *give*, *authority*, and *obedience* are historically on the decline. The graph further shows the decreasing frequency development of the all-animate configuration of the *make*-causative construction. While all four elements show decreasing developments, correlation tests along the lines proposed by Koplenig and Müller-Spitzer (2016), made on the basis of frequency changes between decades, do not indicate a significant correlation between the all-animate variant of the *make*-causative and any of the three lexical elements. This means that the hypothesis is ultimately not supported. While interpersonal authority definitely has become less important to English-speaking societies during the past 150 years, it is not possible to conclude this on the basis of changes in the use of the *make*-causative.

As has been mentioned in earlier sections of this chapter, social change is of course not the only possible explanation for the changes that the *make*-causative has undergone. The fact that the *make*-causative exhibits a collocational profile

with strong preferences for verbs that express experiential responses, emotions, and cognitive processes actually suggests that the notions of subjectification and inter-subjectification (Traugott 1989, 2010) could be fruitfully examined for a more comprehensive account of constructional change (Hilpert 2013; Traugott & Trousdale 2013) in the *make*-causative. To briefly elaborate on this point, Figure 6 shows that the verbs *look*, *seem*, and *appear* undergo frequency increases in the construction over time. A drift towards these epistemic uses of the construction indicates a development that is in line with the tendency that meanings “tend to become increasingly based in the speaker’s subjective belief state/attitude towards the proposition” (Traugott 1989: 35). If diachronic corpus data are used in investigations of social change, it is important to control for processes such as subjectification that may also leave their mark on the linguistic record. In the spirit of Szmrecsanyi (2016), the processes that work together have to be disentangled before we can make claims about any single one of them.

4. Conclusions

This chapter set out to probe how large diachronic corpora can be used for the study of social change. The unprecedented availability of historical databases makes this a very tempting subject of study, and the richness of the data that we currently have at our disposal suggests that fascinating results are in close reach. It has been the aim of this paper to draw together several complicating factors that have been pointed out elsewhere, and to present a case study of the English *make*-causative construction that tries to come to terms with those factors. The main conclusion unfortunately has to be negative. The hypothesis that social change, in the form of waning interpersonal authority, should be visible in the changing use of a grammaticalized construction that serves the purpose of expressing, amongst other things, that kind of authority, could ultimately not be supported. This of course does not mean that it is generally impossible to detect social developments in changing patterns of language use. A phenomenon such as the regular renewal of euphemisms (Pinker 2002) is both socially motivated and clearly visible in the linguistic record. It is also possible to examine meta-linguistic discussions of attempts to actively change language, for instance with the aim of redressing language-inherent gender biases. Hopefully then, the present study has provided some food for thought for endeavors along these lines. In any case, further investigations into the complex interplay of language change and social change are strongly encouraged.

References

Corpora

- COCA = Davies, Mark. 2008. *The Corpus of Contemporary American English* (COCA): 560 million words, 1990–present. <<https://corpus.byu.edu/coca/>>
- COHA = Davies, Mark. 2010. *The Corpus of Historical American English*. <<https://corpus.byu.edu/coha/>>
- NOW = Davies, Mark. 2013. *Corpus of News on the Web* (NOW): 3+ billion words from 20 countries, updated every day. <<https://corpus.byu.edu/now/>>

Other references

- Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257. <https://doi.org/10.1093/lc/8.4.243>
- Flach, Susanne. 2013. Das Ego in der Sprache. *SprachLog*, 12 August 2013, <<http://www.sprachlog.de/2013/08/12/googleology-gets/>> (31 December 2018).
- Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions* [Studies in Corpus Linguistics 39]. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.39>
- Greenfield, Patricia M. 2009. Linking social change and developmental change: Shifting pathways of human development. *Developmental Psychology* 45: 401–418. <https://doi.org/10.1037/a0014726>
- Greenfield, Patricia M. 2013. The changing psychology of culture from 1800 through 2000. *Psychological Science* 24: 1722–1731. <https://doi.org/10.1177/0956797613479387>
- Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2): 243–257. <https://doi.org/10.1515/CLLT.2006.012>
- Hilpert, Martin. 2008. *Germanic Future Constructions. A Usage-based Approach to Language Change* [Constructional Approaches to Language 7]. Amsterdam: John Benjamins. <https://doi.org/10.1075/cal.7>
- Hilpert, Martin. 2013. *Constructional Change in English. Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139004206>
- Hilpert, Martin. 2016. Change in modal meanings: Another look at the shifting collocates of *may*. *Constructions and Frames* 8(1): 66–85. <https://doi.org/10.1075/cf.8.1.o5hil>
- Hilpert, Martin & Correia Saavedra, David. 2017. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory*. (ahead of print). <https://doi.org/10.1515/clt-2017-0009>
- Hilpert, Martin & Gries, Stefan T. 2009. Assessing frequency changes in multi-stage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24(4): 385–401. <https://doi.org/10.1093/lc/fqn012>
- Hilpert, Martin & Perek, Florent. 2015. Meaning change in a petri dish: Constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard* 1(1): 339–350. <https://doi.org/10.1515/lingvan-2015-0013>
- Hopper, Paul J. & Traugott, Elizabeth C. 2003. *Grammaticalization*, 2nd edn. Cambridge: CUP. <https://doi.org/10.1017/CBO9781139165525>

- Kemmer, Suzanne. 2001. Causative constructions and cognitive models: The English *make* causative. In *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). <https://doi.org/10.1371/journal.pone.0150771>
- Liberman, Mark. 2013. The culturomic psychology of urbanization. *Language Log*, 18 August 2013, <<http://languagelog.ldc.upenn.edu/nll/?p=5985>> (31 December 2018).
- Michel, Jean-Baptiste, Shen, Yuan Kui, Presser Aiden, Aviva, Veres, Adrian, Gray, Matthew K., The Google Books Team, Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A., & Lieberman Aiden, Erez. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014): 176–182. <https://doi.org/10.1126/science.1199644>
- McEnery, Tony & Wilson, Andrew. 2001. *Corpus Linguistics: An Introduction*, 2nd edn. Edinburgh: Edinburgh University Press.
- Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics* 54(1): 149–188. <https://doi.org/10.1515/ling-2015-0043>
- Pinker, Steven. 2002. *The Blank Slate: The Modern Denial of Human Nature*. New York, NY: Viking.
- Roberts, Seán & Winters, James. 2013. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLOS ONE* 8(8). <https://doi.org/10.1371/journal.pone.0070902>
- Sagi, Eyal, Kaufmann, Stefan & Clark, Brady. 2011. Tracing semantic change with latent semantic analysis. In *Current Methods in Historical Semantics*, Justyna Robynson & Kathryn Allan (eds), 161–83. Berlin: De Gruyter. <https://doi.org/10.1515/9783110252903.161>
- Szmrecsanyi, Benedikt. 2016. About text frequencies in historical linguistics: Disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory* 12(1): 153–171. <https://doi.org/10.1515/cllt-2015-0068>
- Traugott, Elizabeth C. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 57(1): 33–65. <https://doi.org/10.2307/414841>
- Traugott, Elizabeth C. 2010. Revisiting subjectification and intersubjectification. In *Subjectification, Intersubjectification and Grammaticalization*, Kristin Davidse, Lieven Vandelanotte & Hubert Cuyckens (eds), 27–70. Berlin: De Gruyter Mouton.
- Traugott, Elizabeth C. & Trousdale, Graeme. 2013. *Constructionalization and Constructional Changes*. Oxford: OUP. <https://doi.org/10.1093/acprof:oso/9780199679898.001.0001>
- Turney, Peter D. & Pantel, Patrick. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37: 141–188. <https://doi.org/10.1613/jair.2934>
- Verhagen, Arie. 2000. Interpreting usage: Construing the history of Dutch causal verbs. In *Usage-Based Models of Language*, Michael Barlow & Suzanne Kemmer (eds), 261–286. Stanford, CA: CSLI.
- Zaenen, Annie, Carlette, Jean, Garretson, Gregory, Bresnan, Joan, Koontz-Garboden, Andrew, Nikitina, Tatiana, O'Connor, M. Catherine & Wasow, Tom. 2004. Animacy encoding in English: Why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, Donna Byron & Bonnie Webber (eds), 118–125. East Stroudsburg, PA: Association for Computational Linguistics (ACL). <https://doi.org/10.3115/1608938.1608954>

Changes in society and language

Charting poverty

Gerold Schneider

University of Zurich

This study addresses how societal and linguistic changes can be detected using historical corpora, with the topics of poverty and industrial revolution as a case study, based on large historical corpora, in particular EEBO and CLMET3.0. The results, based on a rich array of state-of-the-art statistical approaches (such as kernel density estimation), show how poverty, industrial revolution, and urbanization are associated through, for instance, the associations of war, religion, family, poverty, and suffering. The study also discusses the importance of data size and cleanness, the temptations of distant reading, and the necessity for validating the discovered patterns in close reading and distant reading in interaction.

1. Introduction

In addition to detecting linguistic change, corpora also provide access to historical texts, allowing insights into developments in society and culture, the history of thought, and the search for specific events in history. The special topic of ICAME 39 – corpus linguistics and the changing society – called for approaches showing how changes in society manifest themselves in corpora, how corpora, including those originally intended for purely linguistic studies, reflect changes in society and events that shaped society.

Finding specific events requires that the event is mentioned in the corpus, which either needs a very large corpus or we will only find events that have left a major impact on history. Finding general trends, however, is also possible with smaller corpora, for example corpora that have been compiled to investigate linguistic changes. Finding such general trends quantitatively has been coined the research area of *culturonomics* by Michel et al. (2011), who used a large sample of books collected by Google, comprising 4% of all books ever printed. This sample is also known as the *Google Books*, and it can be queried online using the Google Books Ngram Viewer.¹ The Ngram Viewer offers a tool for *culturonomics* pilot

1. Available at <<https://books.google.com/ngrams>>.

studies to any interested user as increases, decreases, spikes, and comparison of variant forms can give us preliminary insights both for linguistic and cultural studies.

While Google Books Ngram Viewer is a useful and easily accessible research tool, we go beyond its functionality in various ways in the present chapter. First, we include considerably earlier periods, starting with the *Early English Books Online* (EEBO) collection, whose earliest texts are from 1470. The corpora that we use are described in Section 2. Second, we add several advanced research methods, which allow us, on the one hand, to abstract from individual words to concepts and, on the other hand, free us from needing to start with specific words but detect patterns in the data autonomously.

Methods that detect patterns in the data without any initial guidance, for instance in the form of query terms, are called data-driven approaches (or equivalently corpus-driven). They are frequently used for automated analysis of current media content (Grimmer & Stewart 2013; Wüest, Schneider & Amsler 2014) to detect trends and shifting topics in society, but only rarely in historical linguistics (Hilpert & Gries 2016). We introduce our methods in Section 3.

As area of application, presented in Section 4, we use a case study of societal change: the topics of poverty and industrial revolution. We expect to find that in the 20th century poverty will become a less serious problem. While the industrial revolution has ultimately led to wealth for England and the Western world, the path to it is paved with misery. The debate centrally involves the following two societal questions: In the light of textual evidence,

- Has poverty been a constant threat in Britain, since the middle ages?
- Has the industrial revolution made the English richer or poorer?

In order to assess up to which point the societal changes described in corpora mirror phenomena in society, we are obliged to assess their magnitude by objective, necessarily extralinguistic factors. The extralinguistic debate on poverty is, for example, put by *The Economist* (13 September 2013) as follows: “We know that the Industrial Revolution made Europe rich. But what was it like to live through it? Britain has the most complete historical records” (C.W. 2013). Wages clearly increased only after 1840; from 1700 to 1800 this is debatable. “Access to food was easier for those living in rural areas” (C.W. 2013), compared to frequent misery in the towns. How can we find uncontested objective data assessing the social phenomenon of poverty? While it is generally difficult to find an objective, uncontested gold standard, the case of poverty offers such a touchstone. Here, biological data offers worrying evidence, and holds a real surprise: The mean height of English soldiers decreased between 1730 and 1850; it was only afterwards that, due to improving nutrition and living standards, the mean height increased, as Figures 1 (from C.W., based on Komlos 1998) and 2 (from C.W., based on Hatton & Bray

2010) illustrate. Average life expectancy also reached a low in the 1860s, with only 25 years in Liverpool (Szreter & Mooney 1998, in C.W. 2013).

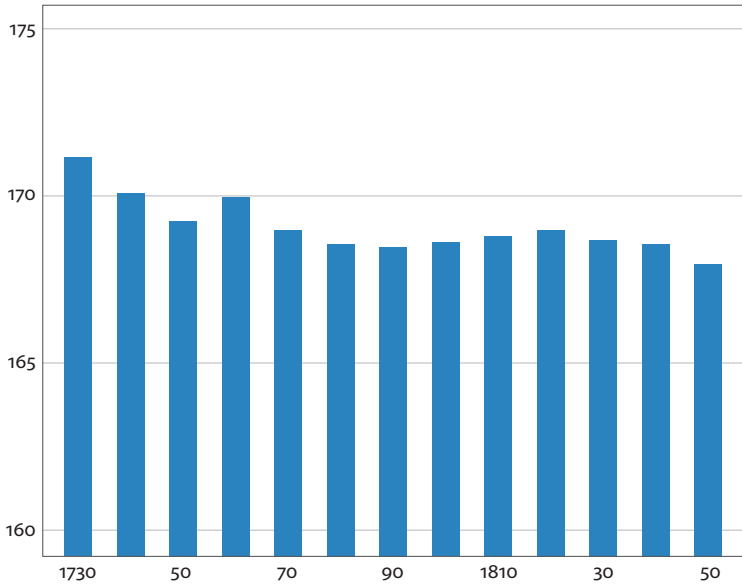


Figure 1. Mean height of English soldiers aged 20–23 years old (cm), from 1730 to 1850
Source: Komlos, 1998

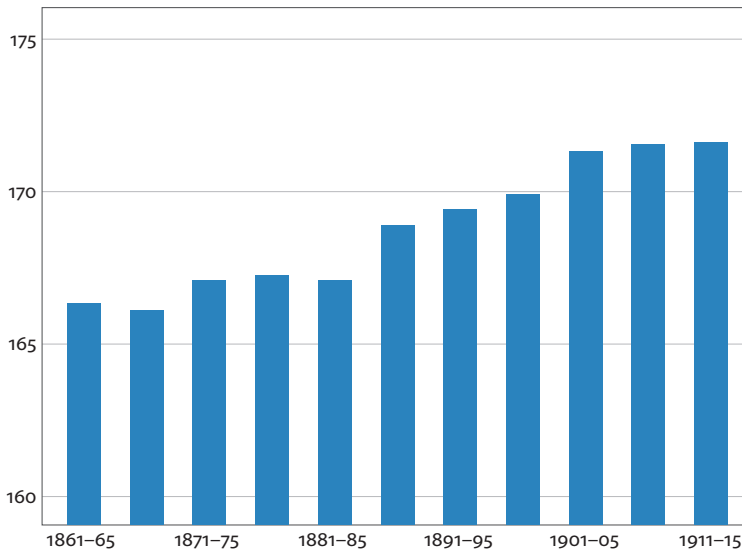


Figure 2. Mean height of men by birth cohorts (cm), from 1861 to 1915
Source: Hatton & Bray, 2010

With this real-world gold standard, we can now assess if poverty is confidently mirrored in the actual corpora under investigation. If so, we can, first, conclude that one can rely on corpora as a social treasure trove. Secondly, we can study attitudes to the extralinguistic phenomenon. How is it framed, is it largely ignored, or is it overrepresented?

We present a case study in which we examine the development of poverty from 1500 to 1920. Poverty and how it can be overcome deals with a core topic in the advance of civilization, human dignity and the success of Western values such as democracy. Will we, for example, find social criticism as in Charles Dickens' works of literature? The role of religion is very important but unclear (e.g. has its importance decreased, what is its attitude?), and the set of query terms for a corpus-based study is very difficult to collect.

As no complete gold standard of social history and the history of thought exists, and as data-driven approaches also partly generate the hypotheses to be investigated, our approach needs to use the extralinguistic proxy of height of soldiers as a second-best proxy, and partly stay exploratory. We give an overview, point out strengths and weaknesses of each approach, and show the complementary nature of the various approaches when applied to our case studies. The main methodological question of our study is: Which method is most appropriate, or do we need a variety of methods? We also discuss the necessity for validating the discovered patterns in close reading and distant reading in interaction.

2. Data and pre-processing

We use a variety of corpora in order to cover a long time period and to investigate generic trends and domain-specific developments.

2.1 The EEBO Collection as sampler corpus

Early English Books Online (EEBO) is not a balanced corpus, but a collection of available books printed between 1470 and 1710, covering about 135,000 books. Over 30,000 of them are now XML-TEI formatted; this part of the EEBO collection is called EEBO-TCP.² The material can be downloaded for free. Despite the handy XML-TEI formatting, the material is difficult to handle. First, errors of tokenization and Optical Character Recognition (OCR) are still frequent, and spelling variants pose a challenge (particularly in data-driven approaches, see Section 3.1). Second, with its size of over

2. 'TCP' stands for Text Creation Partnership. For more information, see <<https://www.manuscriptonline.org/resources/tc/>>.

8 GB, it poses processing challenges. Third, the collection is not balanced (unlike CLMET3.0 or ARCHER), neither by genre nor by period. In the later decades, there is far more material than in the early ones. These issues need to be addressed. We address spelling variants in Section 2.3, and period imbalance in the following.

The fact that EEBO is not genre-balanced is a serious challenge for linguistic investigations that are mainly interested in linguistic change, as shifting genres may overshadow linguistic changes. In our investigation, we are equally interested in changes in content. The fact that certain genres increase or decrease in frequency is part of the equation of changing society, and not a factor that needs to be controlled.

The second problem, that there is far more material in the later than in the early periods, needs to be addressed though, also in order to create a corpus of manageable size. We have created a more balanced random selection, containing 40 million words (to approximately match CLMET3.0 in size). Table 1 provides information on the random sampler that we have created and use in the present study. In the earliest

Table 1. Size of EEBO sampler corpus and percentage of texts used per decade

Decade	Percent of EEBO	Words
1470	100	189,973
1480	100	256,921
1490	100	575,858
1500	100	313,465
1510	100	157,511
1520	100	590,184
1530	80	1,253,102
1540	80	1,458,982
1550	70	1,468,293
1560	70	1,445,497
1570	50	1,674,312
1580	40	1,704,102
1590	20	1,293,852
1600	20	1,328,966
1610	20	1,533,843
1620	20	1,576,422
1630	20	1,418,732
1640	20	3,240,079
1650	20	3,441,874
1660	20	3,047,798
1670	20	3,211,013
1680	20	4,358,888
1690	20	4,364,840
1700	20	336,359
Σ		40,240,866

decades, we have included all texts, and then increasingly fewer, down to 20% in the latest periods. Column 2 gives the percentage of the texts used. Column 3 gives the word count per decade.

2.2 The CLMET3.0 corpus

The *Corpus of Late Modern English texts* version 3.0 (CLMET3.0; De Smet 2005) was compiled by Hendrik de Smet, Hans-Jürgen Diller and Jukka Tyrkkö. It is accessible for free from the University of Leuven.³ CLMET3.0 contains over 34 million words from six coarse genres. It is composed of three periods of 70 years each and covers material from 1710 to 1920. It is a clean and carefully POS-tagged corpus.

2.3 The pre-processing step of spelling normalization

Spelling normalization describes the task of mapping spelling variants to the canonical present-day English (PDE) spelling. We have applied spelling normalization to EEBO with the VARD2 tool (Baron & Rayson 2008). Schneider, Pettersson and Percillier (2017) measured the performance of VARD2 on texts of the same period using the ARCHER corpus. In the 17th century texts, 5.2% of the original tokens are spelled differently from PDE. After the application of spelling normalization, only 1.9% are spelled differently (i.e. the normalization step misses them) or are incorrectly normalized (i.e. they are normalized to a PDE variant, but not the correct one). In the 18th century, 0.8% of the tokens are originally spelled differently, normalization reduces this to 0.3%. The reduction in the 19th century is from 0.6% to 0.4%.

Spelling normalization has an important impact on the number of hits for many queries, and on the performance of subsequent processing steps like tagging, parsing, and the approaches that we use. In topic models, the correlation between topics and non-normalized spelling variants is overestimated if spelling normalization is not used. As a consequence, several of the resulting topics are then dominated by period-specific spelling variants, which means that they tend to over-represent period-specific spelling variants at the cost of representing semantic topics. Topics partly stay stable over time, and partly they change – the element of change is also overestimated if spelling normalization is lacking. If spelling normalization is lacking, period-specific spellings cluster in the conceptual maps – in other words, they tend to deliver clusters of periods instead of clusters of semantically related concepts.

3. Available at <<https://perswww.kuleuven.be/~u0044428/>>.

3. Methods

3.1 Data-based and data-driven approaches

The distinction between corpus-driven and corpus-based approaches is described by Tognini-Bonelli (2001) as follows: in corpus-based approaches, existing hypotheses are tested, while in corpus-driven (also called data-driven) approaches, hypotheses arise from corpus data. Advantages of choosing a data-driven approach are that it needs fewer theoretical assumptions, and that it can bring previously undetected relations and areas of gradience to the surface. A disadvantage is that it is very sensitive to data: "... since the information provided by the corpus is placed centrally and accounted for exhaustively, then there is a risk of error if the corpus turns out to be unrepresentative" (Tognini-Bonelli 2001: 88).

The detection of collocations is a prototypical data-driven approach, but there are many others. Glynn (2010: 10) states:

Many linguists believe that corpus-driven research is restricted to the study of collocations ... Although the study of 'words and the company they keep' lies at the origins of corpus linguistics, methods for studying corpus data have developed much since those times, both in terms of the kinds of questions that are asked and the kinds of techniques used to answer them.

We will present some of the methods used in the present analysis in the remainder of this section. While corpus-driven approaches need fewer theoretical assumptions and are thus a consistent continuation of usage-based linguistics (Bybee 2007) and the quantitative turn (Janda 2013), corpus-driven approaches hardly ever start with a *tabula rasa* as, for instance, Gries (2010) points out. Well-established tools such as word- and sentence-level tokenizers, part-of-speech taggers, often also syntactic frames and alternations (Schneider 2014), are typically relied on. The distinction between corpus-based and corpus-driven is gradient, with corpus-driven approaches offering a much vaster space of results to be reaped from a corpus than a simple *yes* or *no* to a predefined hypothesis, and messier results to wade through and interpret for the researcher, who is at the mercy of the data.

Importantly for semantic investigations, data-driven approaches also allow us to abstract from hand-driven word lists to the automatic detection of concepts, as we discuss in Section 3.3. According to Schwartz and Ungar (2015: 80),

[h]and-driven techniques tend to be more accessible, theory-driven, abstract, and able to handle small datasets, while data-driven tend to be more transparent, capture more connections, and are able to yield unexpected associations.

Data-driven approaches, as they systematically report patterns of difference, are exploratory in nature, allow researchers to find oversights, and are often hypothesis-free, in the sense that the hypotheses are generated by the output of the methods. In many domains of pure science, the hypothesis-driven research method is increasingly supplemented by data-driven approaches. Ananiadou, Kell and Tsujii (2006: 571) state the following:

Systems biology is one of the key examples of a field where the mode of scientific knowledge discovery is shifting from a hypothesis-driven mindset to an integrated holistic mode that combines hypotheses with data ... In the data-rich but hypothesis-poor sciences, including functional genomics and most of biomedicine, the normative hypothesis-driven, deductive scientific method becomes increasingly difficult to sustain because it is unable to deliver advances in knowledge quickly enough.

We apply and compare various methods in this chapter. We use a data-based dictionary approach, and two corpus-driven approaches: topic modelling, and kernel density estimation. Different methods may, on the one hand, perform better or worse for a given task. On the other hand, as additional methods offer different perspectives and different views on the data, it is also possible that they lead to different but compatible and mutually complementary results. In the former case, we can advise which method is preferable and what the pros and cons of each method are. In the latter case, we could conclude that the application of a battery of approaches is essential and summarize what the contributions of the individual methods are.

To analyze corpora by semantic content rather than linguistic differences has a tradition which is at least as well-established as corpus linguistics. The method of *content analysis* has been used in political science and media (e.g. Krippendorff 2004). It has been used in science for a very long time already, long before the computing age, as one of its earliest definitions shows: “analysis of the manifest and latent content of a body of communicated material ... through classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probably effects” (*Webster’s Dictionary of the English Language* 1961).

The fact that the definition explicitly mentions “latent content” is an invitation to use corpus-driven approaches (Schwartz & Ungar 2015). Just like in corpus linguistics, the role of computers and of statistics has increased, allowing large-scale automation. Automated media content analysis “can make possible the previously impossible in political science: the systematic analysis of large-scale text collections without massive funding support” (Grimmer & Stewart 2013: 268).

3.2 Document classification

Document classification is used widely in computational linguistics and in media content analysis. In Grimmer & Stewart (2013) it is given a dominant position. Document classification is used for many tasks, including binary classes of relevant or irrelevant documents for an information retrieval task (see Jurafsky & Martin 2009: Chapter 23.1), authorship identification in forensic linguistics (see Oakes 2014), positive or negative assessment of a political issue in automated content analysis (see Grimmer & Stewart 2013), and binary class of *spam* or *nospam* in most email browsers.

For the most part, words in documents are used as discriminators between the classes given for the classification task, without respecting their sequence or syntactic context. These models are thus called *bag-of-words*. Every word type (above a certain minimal frequency, e.g. $f > 4$) is a feature. This entails that document classification typically uses thousands of features. Each feature in isolation is usually not a good discriminator between the classes, but many ‘weak’ discriminators together achieve very high classification accuracy. Many of the algorithms used are well known in quantitative linguistics, for example logistic regression, but while quantitative linguists use multivariate analysis with perhaps a dozen of features, thousands of features are standard in document classification. The features, for instance words which have the highest weights, can be interpreted as keywords.

The application of document classification, for example predicting periods and interpreting the features, shows mixed success. On the one hand, Yang, Zhu and Tang (2013) detected keywords, while Schneider (2018) interpreted dialect features for comparing Swiss German to German German, and Taavitsainen and Schneider (2019) used document classification to illustrate the movement from scholastic to empirical thinking in the history of medicine. On the other hand, interpreting a feature space of thousands of items is demanding and confusing, and for a semantic investigation the fact that synonyms and related words are completely ignored is a serious disadvantage. We have tested document classification for charting poverty for the present study, by classifying into an early and a late period, but the lists of features were not only very long and thus unwieldy, but also the level of false positives, mirroring corpus coincidences, seemed high. We thus decided to focus on different, more advanced approaches, which systematically map words to concepts (Section 3.3) or show how closely words are related (Section 3.4).

3.3 Topic modelling

In the following subsections, we will use two methods based on distributional semantics. Distributional semantics exploits the so-called Firthian hypothesis, which states that the summed contexts of a word largely define it. The concept is said to go back to Firth who stated that “You shall know a word by the company it keeps” (Firth 1957: 11). More recent formulations are, from a content analysis perspective, “concepts that are closely associated cognitively will also be closely related proximally” (Krippendorff 2004: 206) or, from a computational linguistics perspective “words with similar distributional properties have similar meanings” (Sahlgren 2006: 21).

How are a word’s contexts and its meaning related, then? Quantitative linguistics often exploit distributional information to detect collocations (e.g. Bartsch & Evert 2014; Evert 2008). Sahlgren (2006) shows that the size of the context plays a crucial role. He notes that using adjacent words (or very small observation windows) delivers classical collocations, that is syntagmatic relations, while expanding the window to include increasingly more context delivers results on the paradigmatic axis, i.e. synonyms, antonyms, associations and related words. This is exploited by the hugely successful paradigm of distributional semantics (e.g. Baroni & Lenci 2010).

If the context coincides with the boundaries of the document, we happen to perform document classification. Topic classification achieves high accuracy because single documents are usually dominated by one topic, so that they form a strong unit of discourse, and because important concepts are mentioned repeatedly. Church (2000) observed that the chance of a word to occur in a document radically changes if the same word has occurred before. While a given word w generally has probability $p(w)$, the chance of seeing w a second time if it has occurred once is much closer to $p(w)/2$ than $p(w)^2$ which we would expect if the first and second appearance were independent.

This shift in probability also applies to some degree at the discourse level of topics. Intuitively we know that specific topics are much more likely to generate specific words – this also allows the classification of documents into known classes of topics to perform reliably (see Section 3.2). These insights into the relationship between words, documents and contexts have given birth to the method of topic modelling (Blei 2012), which combines document classification with the strong semantic unity of the discourse of a topic and of a document. Topic models increase the context even beyond the level of the document to the topic level.

Topic modelling optimizes the following equation:

$$p(\text{topic}|\text{document}) * p(\text{word}|\text{topic})$$

for all given documents in a collection. It combines document classification and keyword generation. Document classification can be expressed by the first part

of the above equation, in the Bayesian probability $p(\text{topic}|\text{document})$. The task of document classification is to find the optimal, i.e. most likely class, given the document. The second part of the above formula is the keyword generation probability ($p(\text{word}|\text{topic})$). It expresses that for a given topic we are particularly likely to find certain keywords, as Church (2000) has illustrated at the document level. In document classification (see Section 3.2), the classes are usually given, but not in topic modelling, where the classes are data driven. The classes that emerge from the data often show high semantic coherence, they are therefore called *topics*. *Documents* and *words* are given, *topics* are fitted iteratively starting from a random configuration. The fitting process makes sure that the overall probability of the given documents and words is as high as possible. The user only has to set the number of topics that the algorithm should use. The advantage of topic modelling over dictionary approaches is that instead of thousands of features, only a preselected number of topics needs to be studied by the researcher, and that each topic corresponds to a broad concept.

It can be argued that for well-studied domains, the mapping from words to concepts is available in the form of thesauri, thus no data-driven mapping is needed. Thesauri have been developed carefully and often inspired by painstaking corpus research. A tool which maps words to concepts is the rule-based system *WMatrix*⁴ (Rayson 2008). While this approach performs well on general newspaper texts, specific genres and historical texts often end up with unsatisfactory analyses, as many of the words have undergone strong semantic shifts or have domain-specific readings. Addressing the problem of mapping words to concepts by using contexts in a purely data-driven fashion is a viable and particularly elegant alternative, provided that large amounts of domain data are available.

While *WMatrix* and topic modelling assign words to classes of topics and thus partly solves the mapping from words to concepts, the distances in the distributional space between words are not visible. There are approaches explicitly aiming to assess and display semantic distances in maps of concepts, which we now turn to.

3.4 Conceptual maps

Kernel density estimation, like topic modelling, is also a distributional semantic method. It calculates semantic distances between words by exploiting the observation that semantically related words, for example *horses*, *saddle*, and *ride*, appear mostly in the same locus, both in the real world and in large corpora. If we calculate mutual co-occurrences across texts, we obtain a data-driven measure of similarity.

4. Available at <<http://ucler.lancs.ac.uk/wmatrix/>>.

To illustrate this, we have split the *British National Corpus* (BNC) corpus into 1,000 parts and show the co-occurrence of *horse* and *ride* in Figure 3. We can see that the distribution of the two words shows a partial overlap, but also many differences. The overlap is generally larger than between random words. Particularly in the initial tenth of the BNC, both *horse* and *ride* are quite frequent. As there is an overlap above average, these two words will appear relatively close to each other in the semantic map.

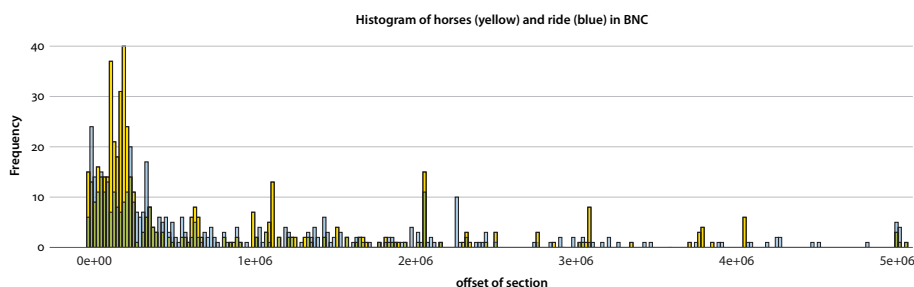


Figure 3. Co-occurrence of *horse* and *ride* in the BNC

Kernel density estimates are kernel functions of these mutual co-occurrences learnt from corpus data. The reason why an approximating kernel function is used instead of the lexical overlap is simply that one can thus obtain smoother results, glossing over data fluctuations. From the co-occurrence matrix, conceptual maps are calculated, which show closely related concepts clustering together in the map, while less closely related words appear increasingly further away from one another on the map. Similar words are plotted close to each other on the arising concept map. We use *textplot* for calculating the kernel densities, and *gephi* for displaying the conceptual maps.

4. Results and discussion

The aim of this study is to investigate the topics of poverty and industrial revolution, based on large historical corpora, in particular EEBO and CLMET 3.0.

4.1 Dictionary-based approach

The main method of research in culturonomics (Section 1) is based on frequencies of selected terms, a method which can already give us insights into many research questions. For example, the saliency of the protestant movement can be traced. The word *protestant* first appears in 1540, stays below 10 occurrences per 1 million

words until 1600, but then rises up to 180 occurrences per 1 million words in 1680. In order to find out mentions of poverty, we search for its most extreme forms: *famine*, *starvation*, and *starving*. The connection between poverty and hunger is obvious, and also stressed by the Food and Agriculture Organisation (FAO) of the United Nations (UN): “Hunger is the most extreme manifestation of poverty and human deprivation.”

We can see slightly decreasing frequencies for the queries with the regular expression given in Figure 4: a word boundary (`\b`), followed by either *povert* or *famine* or *starv*. The regular expression explicitly allows the word to continue, thus *starvation* or *starve* or *starving* will all be matched and returned as results.

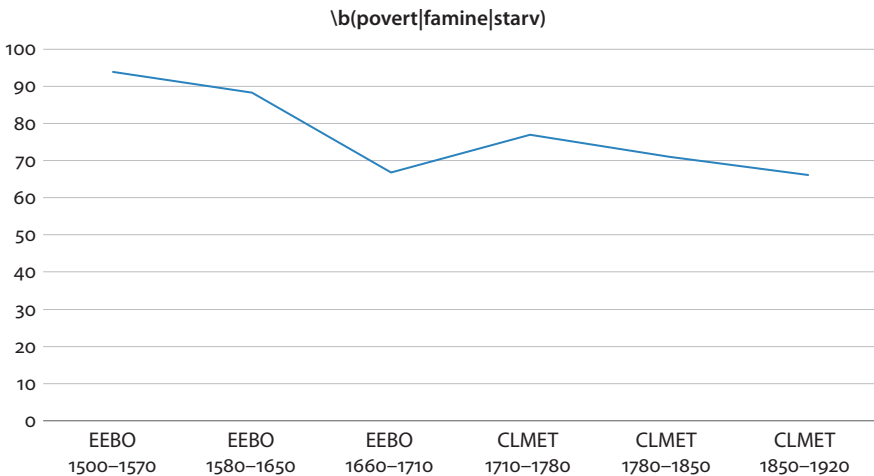


Figure 4. Development of frequency of *famine*, *starvation*, and *starving* from 1500 to 1920, frequency per million words

At a first glance, the issue of poverty seems to stay a constant problem, decreasing slightly, and there is a peak in the period from 1710 to 1780, which at least partly overlaps with the observation of Figure 1 that the mean height of soldiers decreased between 1730 and 1850. The question remains why the period from 1780 to 1850 shows a decrease again. One possible explanation is that after several generations of starvation the topic becomes less salient; it makes ‘fewer headlines’ in today’s terms. It would be desirable to investigate attitudes, associations and the context of the instances. Closer inspection of the query terms also reveals a further shortcoming of relying on single, or a small set of query terms: while *famine* generally decreases, *starv** generally increases.

Similar queries can be made on any other topic of interest. Figure 5 shows the frequency of *trade*, mirroring the expansion and success of the English East India Company and the beginning commerce to North America. Although trade volumes

and the importance of trade for the Gross Domestic Product have increased further (Daudin, O'Rourke & Prados de la Escosura 2008), trade was less represented in writings of the time. Representations in the media are often even better indicators of societal trends than extralinguistic figures.

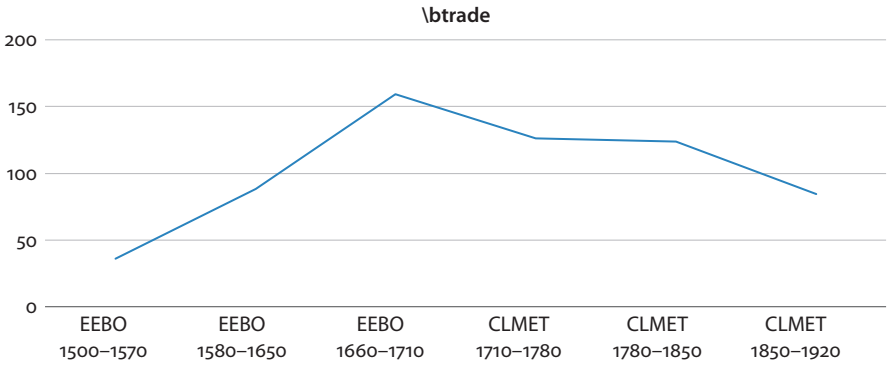


Figure 5. Development of the frequency of *trade* in EEBO. Frequency per million words

Despite this potential advantage, there are several shortcomings of this method. Particularly when using short intervals, there is strong fluctuation due to sparse data, and individual words increase, or fall out of use, or were not known before a certain period. These issues can be addressed by using longer, more carefully compiled lists of synonyms. Based on the OED, we have come up with the following list of expressions referring to poverty:

poor, needful, helpless, wantsome, misease, unwealy, needy, feeble, poorful, mis-eased, indigent, succourless, unwealthy, behove, misterous, miserable, beggarly, starved, threadbare, penurious, fortuneless, wealthless, wantful, necessitous, inopulent, egene, starveling, necessitated, inopious, destitute, want of, pauper, ruined, beggared, impoverished, pennyless, moneyless

While such lists have much higher coverage, thus easing the low recall and sparse data problem, they also tend to overgenerate: for instance, *poor* is often used as a term of sympathy and endearment, *want of* as a general expression of lack.

After spelling normalization and when using the above list of poverty expressions, the differences between the periods are relatively small, as Figure 6 shows. We can mainly observe some fluctuation, the frequency per 1 million words is rising, then falling, rising, and falling again. The variation coefficient (standard deviation / mean) is only 8%. This data suggests that poverty as a problem has not decreased or increased much, or the dictionary method is inappropriate to shed light on changes. Perhaps the attitudes and associations to poverty have changed? The dictionary method cannot help us in answering this question.

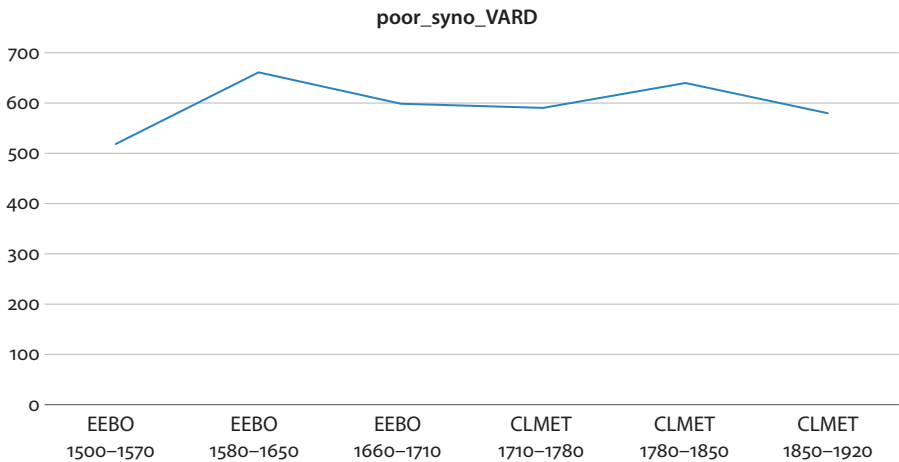


Figure 6. Development of poverty terms in EEBO and CLMET3.0.
Frequency per million words

In order to assess attitudes and associations, we need to resort to more advanced, and more data-driven techniques, as continuing to work with “hand-driven techniques” (Schwartz & Ungar 2015: 80) would become considerably more complicated. We suggest moving to data-driven techniques.

A frequently used method for Information Retrieval and many other Computational Linguistics tasks is document classification (see Section 3.2). While document classification often delivers good results, the results we obtained on the current task were unsatisfactory. In the semantically intricate case of poverty we need to ‘dig deeper’. We thus turn to approaches that aim to step from words to concepts (Section 4.2) or display relations between words (Section 4.3).

4.2 Topic modelling

For our experiments, we use Mallet. Mallet is available as a precompiled tool for all major operating systems. It is written in Java, but can also be called from many other programming languages, for example Python.

Topic modelling should allow us to see in which topics poverty plays a crucial role, how poverty is seen in the eyes of contemporary writers, and which associations it evokes. We hypothesized that we would at least find topics including *religion*, *grief and despair*, and *social and political criticism*, for example the kind we find Charles Dickens’ visionary works of literature.

4.2.1 EEBO early vs. EEBO late

We start by first comparing EEBO early (1500–1599) to EEBO late (1600–1699), the split introduced in the previous subsection. In order to obtain an easily interpretable overview, we use only five topics. The keywords and the contribution to each period are given in Figure 7, sorted from earlier to later importance.

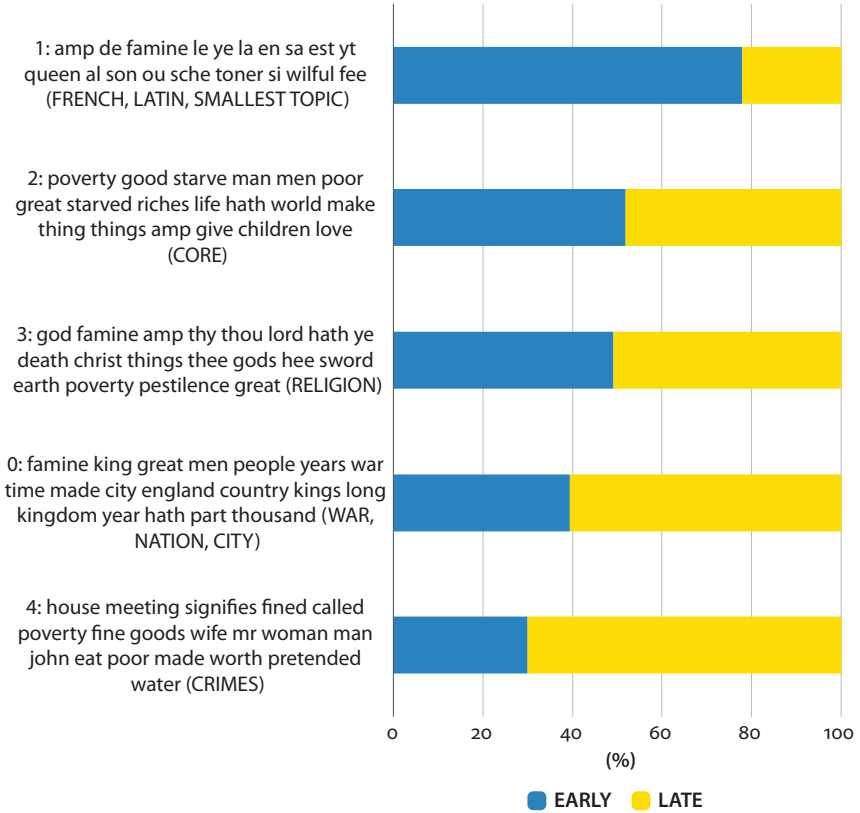


Figure 7. EEBO EARLY vs. EEBO LATE, 5 topics (0–4), sorted from early to late importance. Keywords are on the left, our interpretation in brackets

- Topic 1: the keywords of this topic are mainly Latin and French words. The topic has become less important in the later period, which mainly indicates the decreasing presence of quotes and fixed phrases from Latin. This topic is the smallest topic, only 4.5% of the textual material is associated with it.
- Topic 2: this is the core topic of poverty. This is indicated, first, by the fact that *poverty* and *starve*, two of the most central terms, appear at the very beginning of the keyword list, i.e. they are most important, Second, very frequent,

semantically unspecific words like *man*, *men*, *life*, *world*, and *make* appear centrally. We can also see terms of affection like *love*, which are possibly directed to *children*, showing endearment and parental care. Third, this topic is by far the biggest: 50% of the textual material is assigned to it. This topic is almost equally represented in both periods.

- Topic 3: this topic refers to religion (*god*, *lord*, *christ*). It is also a very important topic, as 27.5% of the textual material is assigned to it. It is equally important in both periods, which indicates that the importance of religion was very strong throughout both periods.
- Topic 0: this topic contains descriptions of *war*, *great men*, the *kingdom of England*. Military actions and praise of heroism have increased in the later period. The 17th century was marked by particularly many wars, both external conflicts including the Anglo-Spanish War, the Anglo-French War, the Anglo-Dutch Wars, and internal conflicts like the English Civil War, the Irish Confederate Wars, and the Monmouth Rebellion. Important sources of the conflicts were the reformation and nationalism. While wars are often glorified in the texts, in practice they brought poverty and misery to the population, particularly in urban areas (Szepter & Mooney 1998).
- Topic 4: this topic has increased strongly in the later period, but it may not appear immediately obvious at first sight. Looking at the documents which are most strongly assigned to this topic reveals that *fine* and *fined* refer to descriptions of *goods* taken illegally. Crimes increase, legal system becomes increasingly important. This is a clear indicator of challenges caused by poverty. The important keyword *house* is highly ambiguous: it sometimes refers to trespassing, to a House of the nobility (who are sometimes victims), and to the House of Commons, so this topic also has a political orientation.

We hypothesized that we would find representations of *religion*, *grief and despair*, as well as *social and political criticism*. Religion appears centrally; grief and despair are also present, for example *death* in topic 3, and *starve* in topic 2, although the topic is expressed less explicitly than expected. We do not see any clear sign of social and political criticism. Topics 0 (WAR, NATION, CITY) and 4 (CRIMES) were not immediately expected.

The strong importance of religion is a mirror of society. The view of religion on poverty partly explains the absence of social criticism and the fact that sadness and despair are less present than expected: poverty is first and foremost a virtue, and in any case a given. Religion and poverty continue to co-exist, and poverty is important for religion. It is typically not portrayed as a problem, but as a virtue, as browsing titles of the important texts reveals:

- 1638: Light from heaven discovering the fountaine opened. Angels acclamations. Churches riches. Rich povertie. In foure treatises
- 1660: The royal and happy poverty or, a meditation on the felicities of an innocent and happy poverty: grounded on the fifth of Matthew, the third verse. And addressed to the late and present sufferers of the times.
- 1685: God's eye on the contrite, or, A discourse shewing that true poverty and contrition of spirit, and trembling at God's Word is the infallible and only way for the obtaining and retaining of divine acceptation.

We only find very few accounts of poverty as a *problem*:

- 1665: Lamentatio civitatis, or, Londons complaint against her children in the countrey shewing her weaknesse, poverty, and desolatenesse ...: as also a brief account how many died in the years 1529 [i.e. 1592], 1603, 1625, 1630, 1636–1637, 1638, 1646–1647, 1648, with this present year 1665
- 1674: 'Tis money makes a man: or, The good-fellows folly. Here in this song good fellow that mayst find, how money makes a man, if thou'rt not blind?

The following quotes from documents that are strongly associated with topic 3 may serve as intermediate summary:

- 1470: "poverty is evil. but evil riches much worse"
- 1470: "Take patiently poverty for the best"
- 1490: "And therefore who that hath riches & loves it becomes pour / & they that have riches & loves poverty is rich ... This grant vs the sweet Ihesus that so much loved poverty"

When working with topic modelling, it is important to read the documents that are assigned to a topic, allowing the interpreter to move between distant reading and close reading (Moretti 2013). For each document, Mallet delivers the percentage with which it assigns a given document to a topic. Particularly those documents that are most strongly assigned to one topic can be considered prototypical of the topic. In order to illustrate such prototypical contents of each topic, we give an excerpt each from the most typical texts (excluding the obvious topic 1, Latin and French expressions):

- Topic 2: To desire in discourse, to hold all arguments, is ridiculous, wanting true judgment; for in all things no man can be exquisite. To have common places to discourse, and to want-variety, is both tedious to the hearers, and shows a shallowness of conceit; therefore it is good to vary, and my speeches with the present occasions, and to have a moderation in all their speeches, especially

in jesting of Religion, State, great persons, weighty and important business, *poverty*, or anything deserving pity;

- Topic 3: Last of all *Gods* love see thy wonderful goodness, which wilt undoubtedly give unto me sanctification & holines for you would not that I should ask for that thing that you wilt not give me. So that I have great cause to lament and rejoyce. To lament, because I am so far from this desire and lamentation, which thy children hav also because of my ignorance, *poverty*, perversity, vn-thankfulness &c. but most of all because thy holy name,
- Topic 0: Hannibal gave diverse assaults to it, but was still repelled with loss, and many sallies they made with variable succes Hannibal mined, and they countermined, so that he was thriven to enclose them, that he might win it by *Famine*.
- Topic 4: Francis Clay was *fined* 5 s. and ten Pound for the pretended *Poverty* of a Preacher. And William Clay was *fined* 10 s. for himself and his Wife, for which he had *Goods* taken from him *worth* 2 l. which the Officers sold for 10 s. Robert Thorocon, the Justice, so *called*, bidding them, Sell the *Goods* for what they could get

4.2.2 Adding CLMET3.0 and increasing the number of topics

As the next step, we include the CLMET3.0 corpus in the comparison. CLMET3.0 includes material from 1710 to 1920. We have also split it into an early and a late period, and used contexts of 1,000 words each around synonyms of poverty (see Section 4.1). We also use a bigger number of topics than in the previous section. Visualizations in the style of Figure 7 are increasingly difficult to interpret. A list of topics and their keywords in which the most important period is added is easier to interpret and to get an overview of. We provide a list of 20 topics in Table 2. They are again sorted from earliest to latest, depending on the period in which the topic was the most important topic compared to the other periods. ID (Column 1) is the arbitrary ID assigned by Mallet. Importance (Column 2) shows the percentage of textual material that is assigned to this topic. We can see, for example, that topic 7, to which we have given a manual interpretation of “General Topic” (Column 3), is the most important topic. In topic models, the most important topic is typically not distinctive and covers general terms (*man*, *make*, *thing*, *world*). The fact that *religion* also appears as a keyword shows that religion plays a very important role in society, throughout all the included periods. Religion also appears in topics 19 (“love of god”), 17 (“religion and sin”), and 12 (“religious institution”). We can observe a movement from adoration to moral issues and the power of the church as institution.

Table 2. Topic model on EEBO and CLMET3.0 with 20 topics, sorted chronologically

ID	Importance	Manual	Strongest period	Mallet keywords
1	0.039	Latin	1 – EEBO EARLY	de est ad queen latin alphabet la si le quote cum ut sed qu quae sunt en quam se
6	0.095	VARD errors	1 – EEBO EARLY	hee bee doe good men man time great things wee thing al make life mind place god end set
19	0.030	Love of god	1 – EEBO EARLY	god man yt good great al men thing love ben lord things ce time se holy fro saith full
7	0.257	General topic	2 – EEBO LATE	man men good great make things reason thing nature world power true part life religion give people made virtue
8	0.068	Politics, Justice	2 – EEBO LATE	parliament law house persons court made time justice day lord act person laws london king lords thereof officers council
9	0.036	Tokenization errors	2 – EEBO LATE	tobe andthe ofgod forthe bythe ofhis thesame thatthe withthe theworld oftheir thatthey fromthe thelord accordingto andthat itis
10	0.152	Art	2 – EEBO LATE	love death life heart soul eyes fair make art great sweet earth day heaven tic made blood gods fear
14	0.108	English history	2 – EEBO LATE	king great prince war england kings queen made majesty people men army duke time france kingdom princes lord english
16	0.064	Drinks and medicine	2 – EEBO LATE	water body good make great time blood drink wine nature made cold put parts hot disease diseases part things
17	0.139	Religion and sin	2 – EEBO LATE	god lord christ sin man life world gods spirit love things holy soul good faith death heaven men mercy
18	0.085	Theatre plays	2 – EEBO LATE	sir man good enter make tic em lady love ha give ll ill isle madam lord poor wife pray
4	0.080	Trade	3 – CLMET EARLY	money country great people trade part england land price year poor pay hundred time labour years greater capital work
11	0.096	Formal letters	3 – CLMET EARLY	mr lord sir great letter day year de duke years george author time dr john london earl house friend

Table 2. (continued)

ID	Importance	Manual	Strongest period	Mallet keywords
13	0.054	Roman empire	3 – CLMET EARLY	emperor roman empire rome city footnote people thousand war years hundred arms romans death public italy age ancient reign
0	0.084	Sea	4 – CLMET LATE	water sea great found ship river day made place country men island time captain miles land small large side
2	0.119	Terms of address	4 – CLMET LATE	mr mrs miss good lady dear mother poor time day father thought house make made young ll man home
3	0.151	Conditio humana	4 – CLMET LATE	nature life human mind man time general society fact character form knowledge world state present idea subject great moral
5	0.215	Body parts	4 – CLMET LATE	man eyes back hand night time face head day long round looked door room moment voice thought stood half
12	0.092	Religious institution	4 – CLMET LATE	church christ god faith religion doctrine holy word true men scripture rome words churches bishop truth gospel christian people
15	0.199	Personal letters	4 – CLMET LATE	mr time sir lady made father heart man friend letter thought dear house make mrs young found good give

For reasons of space, we cannot discuss all remaining topics in detail, but point out a few observations. Topic 8, strongest in EEBO LATE, has an overlap with topic 4 from Section 4.2.1 (CRIMES). Topic 14, English history, is strong in the war-ridden 17th century. Topic 4 (“trade”) is strongest in CLMET EARLY, while seafaring (topic 0) is even stronger in CLMET LATE. Topic 15 (“personal letters”) and 11 (“formal letters”) may be stronger in CLMET3.0 due to the fact that CLMET3.0 is a balanced corpus, while the genre of letters is underrepresented in EEBO, which contains published books. In the latest period, we also see more philosophical texts; Topic 3 contains texts thinking about the *conditio humana* and society.

Coming back to our initial hypotheses, we expected to find representations of *religion, grief and despair*, and *social and political criticism*, but the last of these was largely absent. Inspecting the most prototypical texts, in the following we give an excerpt from the text with the third highest score for topic 3 according to the Topic Model, a philosophical treaty. An example of philosophical speculation, linking physical laws in the solar system to social theory, is given in excerpt (1).

- (1) Every society, while it displays the aggregative process by its increasing mass of population, displays it also by the rise of dense masses on special parts of its area. And in all cases, along with these direct integrations there go the indirect integrations by which parts are made mutually dependent. From this primary re-distribution we were led on to consider the secondary re-distributions, by inquiring how there came to be a formation of parts during the formation of a whole. It turned out that there is habitually a passage from homogeneity to heterogeneity along with the passage from diffusion to concentration. While the matter composing the Solar System has been assuming a denser form, it has changed from unity to variety of distribution. Solidification of the Earth has been accompanied by a progress from comparative uniformity to extreme multiformity. In the course of its advance from a germ to a mass of relatively great bulk, every plant and animal also advances from simplicity to complexity. The increase of a society in numbers and consolidation has for its concomitant an increased heterogeneity both of its political and its industrial organization.
(Herbert Spencer, *First Principles*, 1860)

We still do not easily find social criticism as we had expected, for example the kind we find Charles Dickens' visionary works of literature. We thus increased the number of topics to catch rarer topics. With 50 topics, we obtained a topic that seems to express social criticism. Its keywords are *system, class, fact, population, work, general, society, present, state, pg, time, found, classes, large, great, made, case*.

Is this now the social criticism in the style of Charles Dickens? Inspecting the prototypical documents revealed passages such as (2) and (3).

- (2) It is from the class of females above described, that we naturally look for the highest tone of moral feeling, because they are at the same time removed from the pressing necessities of absolute poverty, and admitted to the intellectual privileges of the great; and thus, while they enjoy every facility in the way of acquiring knowledge, it is their still higher privilege not to be exempt from the domestic duties which call forth the best energies of the female character.

(Sara Stickney Ellis, *The Cult of Domesticity:
A System of Middle-Class Values and Social Duties*, 1838)

- (3) If anything be urgently wanted, it is a plan for preventing the growth of the criminal class; and this probably is not so difficult as it may appear. Of course, till there be a far broader system of public education than now prevails, the criminal population will never want recruits. Nevertheless, even with our present imperfect educational arrangements, something might be done. The criminal class is discovered to be on the whole a narrow class. The practice of living by depredation runs in families, and clings to individuals. The police of any given town could put their hand on almost every person who lives by fraud, theft, and robbery. They could at a day's notice secure nearly every one of them.
(*Chambers's Edinburgh Journal*, No. 419 – Volume 17, 1852)

We conclude that visions of a society that can overcome poverty were rare. In this light, Charles Dickens' works seem all the more visionary, as his contemporaries and even later generations largely had different views.

4.3 Conceptual maps

Topic modelling assigns words to classes of topics, but the distances in the distributional space between words are not visible. While the keywords in a topic are loosely related, it can be desirable to measure and display the semantic distances in maps of concepts. We do so by using kernel density estimation with the tools *texplot* and *gephy* (see Section 3.4).

Figure 8 gives the conceptual map of the contexts of poverty terms from the entire investigated period. In order to keep the complexity of the map manageable, we only keep the 200 most frequent words, after exclusion of stop words such as articles and pronouns. From the list of poverty terms, only *poor* appears among the 200 most frequent words.

There are many relations that we can glean from Figure 8. The word *poor* (towards the left-hand side) appears close to terms of *affection* and *family* (*brother*, *wife*). The close term *friend* is closely related to letter, in which it typically appears as an address. The term *clmetp* is meta-information referring to the CLMET3.0 corpus, thus pointing to later periods. Another close family term, *child*, is conceptually even closer to *mother* and *woman*.

A further term obvious in the poverty domain, namely *miserable*, is most closely related to *life*, *live*, but the *good man* will *pray*, according to the words in the vicinity. *Honour* and *virtue* are important, both on the path to *war* (*city*, *command*, *nation*, *enemies*) and to *heaven*. Death connects *earth* and *heaven*, a main subject of *art*.

Towards the right-hand side of Figure 8, we can see VARD normalization errors (*hee*, *bee*, *doth*, *doe*) closest to *early* (which is meta information, coming from CLMET EARLY), the dominant role of *god* and *holy Christ*, who absolved mankind from *sin*. The CLMET LATE meta term *late* appears closer to the centre of the graph. We see *faith* as the connection to the institution of the *Christian church*, which itself is abused by the *king* or the *prince* to lead *men* to *war*. *War's* promise of *peace* brings on the one hand *power* and *authority* to the *nation*, on the other hand, *blood* is shed and many a *body* of *men* without *fear* will *fall*.

It is remarkable that both the core terms of religion (*god*, *holy*, *christ*) and war (*war*, *king*, *enemies*, *city*, *command*, *peace*) appear quite separated, almost as if outliers, from the core of the graph. *Faith* is the main connection to religion, and *danger* to war.

While conceptual maps as the shown example help one to understand relations, deepen our understanding of history and society, suggest obvious and less obvious connections and relations, they also invite the danger of overinterpretation. At the end of the number crunching process, the researcher is shown a landscape full of

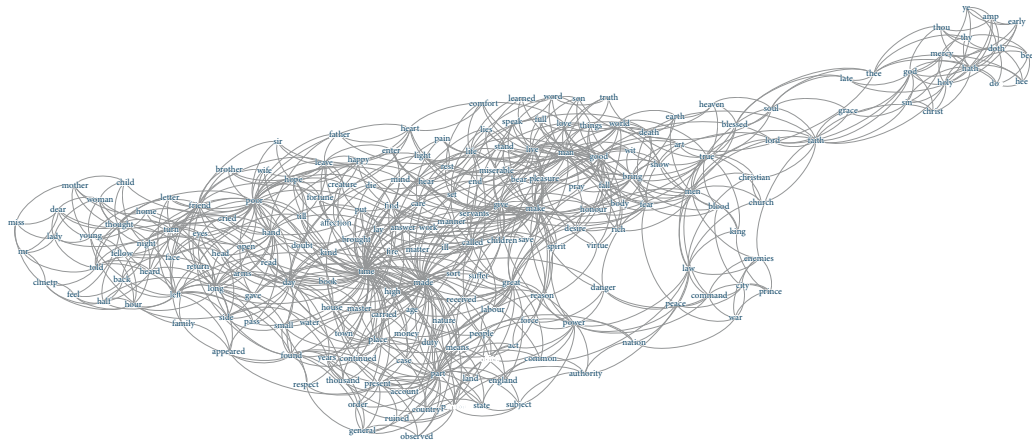


Figure 8. Conceptual map of poverty terms in EEBO and CLMET3.0

inspiration but left with a demanding task of interpretation, akin to the interpretation of a work of literature.

We have created maps of different periods and attempted to compare them to each other, but as the task of interpreting one map is already challenging, comparing different maps to each other is overwhelming. For that task, topic modelling seems more accessible to us.

5. Conclusions

As no gold standard of social history and the history of thought exists, our approach needs to stay exploratory. Due to the nature of our undertaking, conclusions stay tentative. We have investigated frequency and attitudes towards poverty, as expressed in EEBO and the CLMET3.0 corpus. We have compared a data-based dictionary approach and two data-driven approaches: topic modelling and kernel density estimation.

We have seen in Section 4.1 that the dictionary method has failed to deliver interpretable results as frequencies did not change much: its advantage is that it is trivial to apply, its disadvantages are that if the frequency does not change, it cannot provide any answer, and that the list of synonyms is likely to both under- and overgenerate. Still, the method can serve as a first coarse diagnostic step.

In order to assess associations, attitudes and stance towards poverty, we have applied methods using distributional semantics. We have experienced that topic modelling delivers reliable results in Section 4.2. While the core topics and the importance of religion stay stable across the EEBO period, the impact of war and urbanization increases, and even more so the topic covering criminality and the legal system increases, indicating a challenged society. Interacting with the original texts, based on the recommendations that topic modelling offers, reveals detailed insights, for example that poverty is seen rather as religious virtue than a humanitarian problem.

In the CLMET3.0 periods, we experience a shift in the religious topics from religious belief to religious institutions, the increasing importance of trade and seafaring, and more reflections on the human condition in the latest CLMET3.0 period. The quality of topic modelling partly depends on the quality of the input: there are garbage topics (i.e. non-sensical topics, due to data problems) in EEBO, due to tokenization and OCR errors. Nevertheless, the errors tended to cluster in topics of their own, leaving the rest of the findings interpretable.

While topic modelling delivers semantic concepts, and thus partly solves the problem of mapping words to concepts, it cannot answer all questions. Kernel density estimation (Section 4.3) manages better to capture the associations of individual words. It allowed us to see, for example, that *poor* is most often a word

of endearment in the family context, and the connections of war and religion to society. Kernel density estimation reports on the one hand obvious connections, such as the strong topic of religion, but also many more, such as the fact that wars are fought in the name of or under the pretext of peace. The method offers a tremendous source of inspiration, but also bears a danger of overinterpretation.

Coming back to the methodological questions of our study (which method is more appropriate, or do we need a variety of methods), we conclude that there are clear differences: dictionary methods did not deliver interpretable results for our task, document classification did not work well, while data-driven distributional semantics led to the hoped insights. Topic modelling and kernel density estimation delivered different but compatible results, offering complementary perspectives on the data. We were also intrigued to see that at the end of the number crunching, the most demanding task is the interpretation. We have learned a lot about society, history and automated media content analysis.

Coming back to our initial societal questions, we found that poverty remained a problem until the 20th century, but that attitudes changed. We saw the central role of religion confirmed, but also observed a shift from adoration to the power of the institution of the Christian church. We partly found grief and despair, but more frequently contexts of affection and family terms. Social criticism was almost completely absent, and instead we found expressions of contempt and calls for tougher legal action in the CLMET3.0 period.

References

Corpora and software

- CLMET3.0 = *Corpus of Late Modern English Texts*, version 3.0. De Smet, Hendrik, Diller, Hans-Jürgen & Tyrkkö, Jukka (comps). <<https://perswww.kuleuven.be/~u0044428/>>
 EEBO = *Early English Books Online*. Davies, Mark (comp.). <<https://eebo.chadwyck.com/home>>
 Gephi = <<https://gephi.org>>
 Mallet = Machine Learning for Language Toolkit. <<http://mallet.cs.umass.edu/topics.php>>
 Textplot = <<http://dclure.org/tutorials/textplot-refresh/>>
 VARD2 = Baron, Alistair and Rayson, Paul. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, 22 May 2008.

Other references

- Ananiadou, Sophia, Kell, Douglas B. & Tsujii, Jun-ichi. 2006. Text mining and its potential applications in systems biology. *Trends in Biotechnology* 24(12): 571–579.
<https://doi.org/10.1016/j.tibtech.2006.10.002>
 Baroni, Marco & Lenci, Alessandro. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics* 36(4): 673–721.
https://doi.org/10.1162/coli_a_00016

- Bartsch, Sabine & Evert, Stefan. 2014. Towards a Firthian notion of collocation. In *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern* [OPAL - Online publizierte Arbeiten zur Linguistik 2/2014], Andrea Abel & Lothar Leimnitz (eds), 48–61. Mannheim: Institut für Deutsche Sprache.
- Blei, David. 2012. Probabilistic topic models. *Communications of the ACM* 55(4): 77–84.
<https://doi.org/10.1145/2133806.2133826>
- Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford: OUP.
<https://doi.org/10.1093/acprof:oso/9780195301571.001.0001>
- Church, Kenneth. 2000. Empirical estimates of adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2 . *Proceedings of the 17th Conference on Computational Linguistics*, 180–186. Stroudsburg, PA: Association for Computational Linguistics.
- C. W. 2013. Did living standards improve during the Industrial Revolution? *The Economist*, September 13, 2013. <<https://www.economist.com/free-exchange/2013/09/13/did-living-standards-improve-during-the-industrial-revolution>> (30 December 2018).
- Daudin, Guillaume, O'Rourke, Kevin H., & Prados de la Escosura, Leandro. 2008. Trade and empire, 1700–1870. Technical Report # 2008–24, OFCE: Centre de recherche en économie et sciences po. <<https://www.ofce.sciences-po.fr/pdf/dtravail/WP2008-24.pdf>> (30 December 2018).
- De Smet, Hendrik. 2005. A corpus of Late Modern English. *ICAME Journal* 29: 69–82.
- Evert, Stefan. 2008. Corpora and collocations. In *Corpus Linguistics. An International Handbook*, Anke Lüdeling & Merja Kytö (eds), 1212–1248. Berlin: Mouton De Gruyter.
- Food and Agriculture Organisation of the United Nations. November 2003. Anti-hunger Programme. *A twin-track approach to hunger reduction: priorities for national and international action*. <<http://www.fao.org/3/J0563E/j0563e02.htm>>.
- Firth, Rupert. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis* [Special volume of the Philological Society], Rupert Firth (ed.), 1–32. Oxford: Blackwell.
- Glynn, Dylan. 2010. Corpus-driven cognitive semantics. Introduction to the field. In *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* [Cognitive Linguistics Research 46], Dylan Glynn & Kerstin Fischer (eds), 1–42. Berlin: Mouton de Gruyter.
- Gries, Stefan T. 2010. Corpus linguistics and theoretical linguistics: a love-hate relationship? Not necessarily... *International Journal of Corpus Linguistics* 15(3): 327–343.
<https://doi.org/10.1075/ijcl.15.3.02gri>
- Grimmer, Justin & Stewart, Brandon. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.
<https://doi.org/10.1093/pan/mpso28>
- Hatton, Timothy J. & Bray, Bernice E. 2010. Long run trends in the heights of European men, 19th–20th centuries. *Economics & Human Biology* 8(3): 405–413.
<https://doi.org/10.1016/j.ehb.2010.03.001>
- Hilpert, Martin & Gries, Stefan T. 2016. Quantitative approaches to diachronic corpus linguistics. In *The Cambridge Handbook of English Historical Linguistics*, Merja Kytö & Päivi Pahta (eds), 36–53. Cambridge: CUP.
- Janda, Linda A. 2013. *Cognitive Linguistics: The Quantitative Turn*. Berlin: Mouton de Gruyter.
- Jurafsky, Daniel & Martin, James H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.
- Komlos, John. 1998. Shrinking in a growing economy? The mystery of physical stature during the industrial revolution. *Journal of Economic History* 58: 779–802.
<https://doi.org/10.1017/S0022050700021161>

- Krippendorff, Klaus. 2004. *Content Analysis*, 2nd edn. London: Sage.
- Michel, Jean-Baptiste, Shen, Yuan Kui, Presser Aiden, Aviva, Veres, Adrian, Gray, Matthew K., The Google Books Team, Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A. & Lieberman Aiden, Erez. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014): 176–182.
<https://doi.org/10.1126/science.1199644>
- Moretti, Franco. 2013. *Distant Reading*. London: Verso.
- Oakes, Michael P. 2014. *Literary Detective Work on the Computer* [Natural Language Processing 12]. Amsterdam: John Benjamins.
- Oxford English Dictionary*. 2010. 3rd edn. Oxford: OUP.
- Rayson, Paul. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4): 519–549. <https://doi.org/10.1075/ijcl.13.4.06ray>
- Sahlgren, Magnus. 2006. The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces. PhD dissertation, Stockholm University.
- Schneider, Gerold. 2014. *Applying Computational Linguistics and Language Models: From Descriptive Linguistics to Text Mining and Psycholinguistics*. Cumulative Habilitation, University of Zurich.
- Schneider, Gerold. 2018. Differences between Swiss High German and German High German via data-driven methods. In *Proceedings of SwissText 2018*, Mark Cieliebak, Don Tuggener & Fernando Benites (eds), 6–16. <<http://ceur-ws.org/Vol-2226/>> (30 December 2018).
- Schneider, Gerold, Pettersson, Eva & Percillier, Michael. 2017. Comparing rule-based and SMT-based spelling normalisation for English historical texts. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. <<http://www.ep.liu.se/ecp/133/008/ecp17133008.pdf>> (30 December 2018).
- Schwartz, H. Andrew & Ungar, Lyle H. 2015. Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science* 659(1): 78–94. <https://doi.org/10.1177/0002716215569197>
- Szreter, Simon & Mooney, Graham. 1998. Urbanization, mortality, and the standard of living debate: new estimates of the expectation of life at birth in nineteenth-century British cities. *Economic History Review* 51(1): 84–112.
- Taavitsainen, Irma & Schneider, Gerold. 2019. Scholastic argumentation in Early English medical writing and its afterlife: New corpus evidence. In *From Data to Evidence in English Language Research* [Language and Computers 83], Carla Suhr, Terttu Nevalainen & Irma Taavitsainen (eds), 191–221. Leiden: Brill. https://doi.org/10.1163/9789004390652_010
- Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.
- Webster's Dictionary of the English Language*. 1961. Springfield, MA: Merriam-Webster Incorporated.
- Wüest, Bruno, Schneider, Gerold & Amsler, Michael. 2014. Measuring the public accountability of new modes of governance. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, Maryland*, 38–43. Stroudsburg, PA: Association for Computational Linguistics.
- Yang, Li-gong, Zhu, Jian & Tang, Shi-ping. 2013. Keywords extraction based on text classification. *Advanced Materials Research* 765–767: 1604–1609.
<https://doi.org/10.4028/www.scientific.net/AMR.765-767.1604>

Finding evidence for a changing society

A collocational study of medical discourse in 1500–1800

Maura Ratia
University of Helsinki

This chapter examines, with the help of collocation analysis, how patients were viewed in medical texts from 1500 to 1800. Previous studies have suggested that this period witnessed considerable changes in society. The field of medicine also underwent major developments during this time, but linguistic analyses have been lacking. Two corpora were used in the study: the *Corpus of Early Modern English Medical Texts* and the *Corpus of Late Modern English Medical Texts*, together totaling over 4.2 million words. The results indicate a development from the patient as an object of various treatments and cures in the early modern period to patient as *experiencer* in the late modern period. The growing importance of hospitals and public health in the latter era also emerges from the results.

1. Introduction

In terms of scientific advancement and medical writing, the period from 1500 to 1800 can be described as transformational. The early modern period brought about new scientific methods of observation and experiment as well as new genres of writing, and, in a similar vein, the spirit of Enlightenment of the 18th century saw groundbreaking advances in medicine, such as smallpox inoculation, accompanied by a better understanding of diseases. In the latter period, novel cultural ideals also developed. This is manifested in, for example, the establishment of philanthropic societies and especially the emergence of “polite society”. Politeness, benevolence and compassion were among the traits that were expected of “medico-gentility” of the time (Brown 2011: 74), and the newfound concerns about the humane treatment of patients entered the medical discourse in tandem with the new scientific innovations.¹

1. I would like to express my gratitude to the anonymous reviewers for their critical yet helpful comments on earlier drafts of this chapter. I also wish to thank my colleagues, Tanja Säily and Turo Vartiainen, for their advice on my poster presentation at the ICAME39 Conference in Tampere 2019, which acted as the basis for this study.

In the present study, I explore how the new and changing ideals, specifically those of polite society of the 18th century, are represented in medical discourse. Earlier studies have noted a transition in polite discourse from the cultural keywords of *courtesy* and *civility*, dating back to Late Middle Ages and Renaissance, to *politeness* in the 18th century (Nevalainen & Tissari 2010: 134). In Late Modern England, the privileged and strong institutional status of doctors of the Stuart era had been weakened (Porter 1986: 287). At the same time, medical literature, especially in the form of popular journals, and familiarity with medicine was “integral to the public role of the well-informed, public-spirited, and responsible layman” (Porter 1985: 163).

As previous studies suggest that from the late 17th century onwards significant changes took place in society (see e.g. Klein 2002), this chapter examines whether and how this can be evidenced in medical writing and, in particular, in the overall attitude of physicians towards their patients. Most of the authors were medical practitioners themselves, although surgeons and medical writers without formal medical training also contributed to the discussion.

One of the hypotheses for this study is that collocation analysis can provide new information about how patients were seen in the early modern and late modern periods and how these ideas were realized linguistically (see e.g. McEnery & Hardie 2013). Another hypothesis is that in the later period we see more expressions of empathy towards patients, which would conform with the cultural ideas of the time, emphasizing compassion towards fellow human beings. Collocation analysis as a method can reveal conventionalized expressions that might not be noticed otherwise. The frequency of these expressions makes them possible carriers of cultural and domain-specific meanings (see Pollach 2012: 270). The words used to extract collocates in the material were *patient* and *patients*.

Two corpora were used as material: EMENT, the *Corpus of Early Modern English Medical Texts*, totaling over 2 million words, and LMENT, the *Corpus of Late Modern English Medical Texts* with 2.2 million words. LMENT as a new tool allows for studies of medical writing in a long diachrony, which were previously difficult to conduct due to a lack of corpus material suitable for linguistic study.

The results of the study show a shift in the 18th century from the patient as an object of treatments to the patient as *experiencer*. The effort to treat patients more humanely emerges especially in the use of negative vocabulary bringing patients' experiences to the fore. Specific medicines and cures are discussed less in this later period than in early modern texts. The analysis of the plural node *patients*, in turn, illustrates the general 18th-century trend of improvement in medicine including the development of voluntary hospitals and medical legislation, and the way these topics are dealt with in medical writing (Lehto 2019: 315; Porter 1991: 65).

2. Background

The term *collocation* fundamentally refers to “the pervasive regular patterning of co-occurrence among words in corpus data” (McEnery & Hardie 2013: 742) or to non-sequential, habitual patterns of co-occurrence (Stubbs 1995: 23). However, the specifics of understanding what collocation analysis actually determines or how collocations are calculated vary (see Gries 2013). Collocation analysis, as well as other corpus-linguistic methods, demands meticulous planning because the settings greatly affect the results. For instance, collocation strength can be measured in different ways, for example, with an MI measure and a t-score. The MI is more prone to find rare collocations whereas the t-score finds frequent collocations (Kilgariff 1996: 3); some studies have also combined these two measures (see Nilsson Björkenstam, Gustafson Capková & Wirén 2014). The use of log-likelihood measure has also been advocated and t-score is used less nowadays (see Evert 2005; Pollach 2012: 270; also Stubbs 1995: 33–34 for criticism of t-score).

Collocations are commonly assessed with three criteria: distance, frequency and exclusivity (Brezina, McEnery & Wattam 2015: 140). Collocational distance to the node word is termed the *collocation window* or *span*, commonly ranging from as little as one word to four or five on both sides of the node (see also discussion in Barnbrook, Mason & Krishnamurthy 2013: 166–167). When the frequency of collocation is considered, the focus of the study is essential, that is, whether it has to do with more general or immediate associations of a given word. With historical material, one should not limit the analysis too much beforehand. If the frequency floors are set too high, for instance to the often suggested 10, the analysis can be difficult to conduct because that might not provide enough material for analysis. Exclusivity of collocation can also be determined, indicating whether the link between the node and the collocate is exclusive. However, this criterion was not considered in the present chapter because it would have reduced the number of collocates too much in the datasets.

One of the terms that have been used in connection with collocation analysis is *semantic prosody* referring to the idea of words creating some kind of transference of meaning to the habitual collocates surrounding them, that is the company they keep (Louw 1993: 158–159; see Stubbs 1995). This tendency of words to denote positive or negative prosodies or associations and to undergo semantic changes through time is relevant for a diachronic study like the present one.

In addition to examining single words, the presence of *lexical bundles* also needs to be considered with some of the collocates. The term *lexical bundle* refers to kinds of extended collocations of sequences of three or four words, which have a statistical tendency to co-occur in the given material (Biber & Conrad 1999: 183–184). If compared to idioms, lexical bundles are more frequent but less salient and they can often be structurally incomplete, which makes it harder to spot them without the assistance of statistical tools.

3. Materials and method

In order to trace larger trends in historical data, corpus-linguistic methods are a welcome addition to the more traditional approaches of, for example, close reading of texts and studying etymologies in dictionaries. Comparisons between the two corpora, EMENT and LMENT, are enabled by the fact that both corpora were compiled according to the same principles. EMENT and LMENT² are components in the larger *Corpus of Early English Medical Writing* (CEEM), comprising three parts of which the first one, the *Corpus of Middle English Medical Texts* (MEMT), is not included in this study. In Figure 1, the division into categories in the respective time periods and corpora is shown. The categories reflect the topic of the text or the field of medicine, determined with the help of reference works, primary sources and online resources, and in collaboration with historians of medicine (Pahta & Taavitsainen 2010: 5). Because of the multifarious nature of medical writing in 1400–1800 consisting of various genres, general topics of the texts “provide a solid basis for longer diachronic assessments of this language variety” (Taavitsainen et al. 2014: 139). For instance, within the category of SPECIFIC TOPICS, epidemic diseases, such as the plague and smallpox, continued to be discussed, but new topics also surfaced including dentistry and mental health (Ratia 2019: 243–245).

The material for this study comes from the period of 1500–1800. The categories remain rather constant: one notable change lies in the specialization of scientific journals in the 18th century. In the EMENT corpus, a category titled *Philosophical Transactions* (1666–1700), representing journals, is divided further into three parts in the LMENT corpus. It constitutes as much as 24% of the text population, which contrasts to the mere 8% in the earlier EMENT corpus, underlining the significance of medical journals in the 18th-century medicine (Hiltunen 2010: 127; Taavitsainen et al. 2014: 138–139). In addition to *Philosophical Transactions*, which became more established and theoretical in this period (Atkinson 1992: 340), the 18th century saw the first academic journals dedicated to medicine: the learned *Edinburgh Medical Journal* (EMJ) consisting of transactions of medical societies, and the *Gentleman's Magazine* (GM), a popular journal containing medical topics (Porter 1985: 141–143, 163–164). Another change in the constellation of categories is the birth of a new category, namely institutional texts, which comprises 10% of the corpus material. Works in this category were not yet part of the national movement of public health but they were rather pioneering works by individual authors (Lehto 2019: 307; Porter 1991: 65). The topics discussed in these texts range, for instance, from hospitals and legislation to improving public health and hygiene. The distribution of medical journals and institutional texts in the LMENT corpus is uneven as both EMJ and GM were primarily published in the latter half of the 18th

2. A pilot version of the LMENT corpus was used for the study.

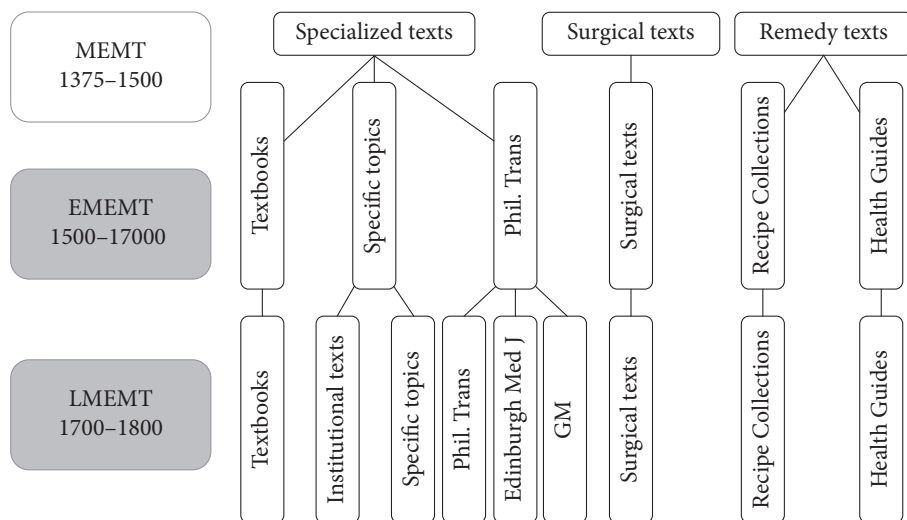


Figure 1. Categories of medical writing in the *Corpus of Early English Medical Writing* (based on Taavitsainen et al. 2014: 140)

century: EMJ in 1747–1796 and GM in 1731–1798. Similarly, institutional texts became more common during this same period because of the growing importance of hygiene and statistical methods (Lehto 2019: 315).

In EMENT, normalized text-files incorporated in the corpus were used for analysis to minimize the effects of spelling variation on the results because Early Modern English was still characterized by the lack of standardized spellings, although, by the middle of the 17th century, the process towards more systematic spellings had already become evident (see Lehto et al. 2010). The normalization was automatically processed with a manually trained version of the VARD software (Variant Detector, version 2.3, Baron & Rayson 2009).³

In this study, AntConc (Anthony 2014) was used to measure collocations, and the default measure MI (Mutual Information) was employed to measure collocation strength. The ability of the MI measure to trace less frequent associations of words was desirable in order to fully examine the potential of collocations to convey various attitudes, even if these associations are not necessarily “generalisable to other samples of language” (Durrant & Doherty 2010: 131). In the EMENT and LMENT corpora, the node word *patient/s* is a relatively frequent content word in the register of medical writing; The singular *patient* occurs 1,097 times in the EMENT, of which 91 occurrences are in the possessive form *patient’s*. The plural *patients*

3. The normalized texts in EMENT contain less than 5% of unknown variants in comparison to over 18% in the non-standardized texts (Lehto et al. 2010: 287–289).

occurs 145 times. Content words are, nevertheless, more infrequent than some function words such as pronouns, which had to be taken into account. In order to control this sensitivity of MI to rarely occurring items, frequency floors were used. In this study, the frequency floor was lowered to five, to give enough variation in the results; this same threshold has been suggested by Church and Hanks (1990; cited in Kilgariff 1996: 3).

As within any dataset, especially a historical one, texts can be quite heterogeneous and a single text can stand out from the rest whether it is due to idiolectal, dialectal, or sociolectal variation. To reduce the possible overpowering influence of one author or one text over others in the results, the collocates found in fewer than three different texts were excluded.⁴

Commonly the distance between the node word and its collocates is set to four or five words to the left and the right of the node (see Brezina et al. 2015: 140). The distance of five, the default setting in AntConc, was applied in this study so that as much information as possible would be retained, but this also demanded scrutiny to determine potential incorrect collocates. Using a smaller scope would provide more exact results, but the need to qualitatively examine collocates in their context still remains. When the data was analyzed, some occurrences of the collocates were clearly incorrect. For instance, *medicines* followed the node *patients*, which was the last item on the page, as part of a heading on the following page “The composition of *medicines* for wounds” (Baker, *Oleum magistrale*, 1574). In this particular case, having a smaller span of words instead of five left and right would have helped, but it is just as possible that the heading could have started with the word “Medicines”, in which case even the strictest search criteria would not have eliminated this problem.

Each concordance line had to be checked to confirm the meaning of the word in that specific context. For instance, the strongest collocate for the node word *patient* in the EMENT corpus was “agent”, which had to be excluded from the analysis as in all examples the denotation of *patient* in phrases such as “agent and patient”, was not of someone receiving medical treatment but of a passive object of action, which is the other, more infrequent, meaning of the word. In the same vein, collocates with ambiguous or multiple referents were omitted.⁵ In this study, lemmatization was not used because of the frequency of the examined collocates was not so high, and the word forms themselves were considered to be important.

4. The following collocates were excluded for the plural node *patients*: In EMENT *whether*, which was found in only two texts, and in LMENT *incurable*, similarly found in two texts, and *affluent*, *payments*, and *maintenance*, which only occurred in an anonymous letter from 1788 titled *Letter to York lunatic asylum*.

5. Two such ambiguous collocates were excluded from EMENT results: the collocate *leaving* with the singular node *patient* and the collocates *both*, *were*, *had*, and *what* with the plural node *patients*.

4. Results

In the analysis of collocates, the ten strongest collocates for the word *patient/s* were examined in the EMENT and LMENT corpora. The analysis is divided into two parts: first the collocates are analyzed in chronological order in EMENT and then in LMENT, to make it easier to draw diachronic comparisons between the early modern and late modern periods. Normalized corpus texts were used in the quantitative analysis of collocates but in the micro-analysis of collocates original spellings are retained.

4.1 The Corpus of Early Modern English Medical Texts (1500–1700)

When collocates are extracted from the EMENT corpus, the word *strength* emerges as the strongest collocate for the word *patient* (see Table 1). Almost all references are to “the strength of the patient” or “the Patient’s strength”, having to do with administering proper doses of medication or performing other treatments to the patient, which is one of the core functions of medical texts.

Table 1. The strongest collocates for the singular node *patient* (on the left) and the plural *patients* (on the right) in EMENT

EMENT							
Collocates	Coll. freq.	Texts	MI	Collocates	Coll. freq.	Texts	MI
<i>Patient</i> (1,097)				<i>Patients</i> (145)			
<i>strength</i> (580)	61	30	7.425	<i>strength</i> (580)	10	4	7.735
<i>smoke</i> (74)	5	4	6.786	<i>their</i> (5,035)	35	21	6.425
<i>let</i> (2,546)	135	46	6.437	<i>our</i> (1,831)	9	4	5.925
<i>purged</i> (170)	9	6	6.434	<i>my</i> (1,531)	7	4	5.820
<i>lie</i> (180)	9	6	6.352	<i>medicines</i> (1,102)	5	5	5.809
<i>recovered</i> (103)	5	4	6.309	<i>have</i> (6,657)	23	14	5.416
<i>recover</i> (105)	5	4	6.281	<i>drink</i> (1,865)	5	5	5.050
<i>sweat</i> (443)	21	15	6.275	<i>such</i> (3,801)	10	7	5.023
<i>ease</i> (256)	12	11	6.259	<i>let</i> (2,546)	6	5	4.864
<i>give</i> (1,410)	63	32	6.190	<i>his</i> (5,627)	12	9	4.720

In addition to the collocate *strength*, another collocate, *let*, is used with *patient*. Almost all examples are of the type “let the patient”, which is a frequent lexical bundle in the material (114 occurrences). In (1), advice is given for “Fits of the Mother, and Vapours from the Spleen” (node word and collocates are italicized in all examples).

- (1) *LET the Patient* (having regard to the direction of the general Observations) begin with one, or with two Spoonfuls of this Drink, and so continue and raise the Dose, according to what the *Patients strength* may admit, remembring that the greatest Dose prescribed is not to exceed three Spoonfuls, and let the proper Dose be taken at any time whatsoever, when the Fit is found to be coming upon him or her.
(Daffy, *Directions for taking Elixir Salutis*, 1700: 6–7)

Imperatives are typically employed in instructive passages across genres in the EMEMT corpus: from general and specific treatises to regimens and recipe collections (Pahta & Ratia 2010: 82; Taavitsainen & Tyrkkö 2010: 72). The recurrent lexical bundle “let the patient” can also be found with the collocate *smoke*, as in (2).

- (2) [...] then, by the litle hole aboue, *let the patient* take the *smoke* into his eare, by the space of halfe an hower.
(Ruscelli, *Secretes of Alexis of Piemovnt*, 1562: f. B4v)

In these examples, the word *patient* functions as the object whereas the physician is the active force behind action – not specifically mentioned as such, but implied in the context.

The collocate *purged* (as in (3) and (4)) occurs most commonly with different variations of “be purged” constructions, with the *patient* again as the object of action. Purging and blood-letting were at the time common methods of cure to restore a healthy balance of humors and temperament in a patient’s body.

- (3) And in this case, after the Letting of Blood, the *Patient* may be gently *purgd* on the thirteenth, or any other succeeding day, but not before; [...]
(Sydenham, *Compleat method of curing*, 1695: 37)
- (4) [...] let a plaister of mustard-seede bee layed to the nape of the necke; wee will not vse these drawing medicines before the *patient* bee 3. or 4. times *purged*.
(Bruele, *Praxis medicinae*, 1632: 18, orig. 26 [sic])

In (4), the presence of physicians is realized with the first-person plural pronoun (*we*). The reference is general in nature and has a vocative function implicated by the use of the modal verb *will* (see “workshop *we*” in Wales 1996: 65–66).

The verb *lie* is used in a similar fashion with imperative verb forms, the most common usage being “let the patient lie” as in (5) describing a “treue medicine for deafness of ears”.

- (5) Take Beten & Horhou~de and stamp them two in a mortar / & wring out the juice / and let the *patient lie* on the side / & power it into his eere / and let him rest.
(Anon., *Treasure of pore men*, 1526: f. 13v)

The context of situation with the lexical bundle “let the patient” conveys passive action on the part of the patient and gives instruction to the readers, mostly consisting

of other physicians but, in the course of the early modern period, to other men of learning and even to lay people of the middling classes. Only one of the occurrences of the collocate *lie* expresses necessity rather than recommendation with the modal verb *must*. The patient is also emphasized as the subject of the clause, perhaps because of the context of a very painful affliction, that is, the colic (as in (6)).

- (6) Civet is most excellent in this Disease: for the quantity of a Pease, applied to the Navil, and a hot Loaf out of the Oven clapt over it, presently easeth the pain: the *Patient* must *ly* on his Belly upon the Bread before it be cold.

(Porta, *Natural magick*, 1658: 222)

When the collocates *recovered/recover* are used, attention is focused on the patient as the subject. Example (7) is part of a case narrative or report, which were essential discourse components in early modern scientific writing used to give credibility to arguments (Gross, Harmon & Reidy 2002; Taavitsainen 2011: 93).

- (7) His Friends then urged a second Transfusion, which being at last performed the next morning, the *Patient* indeed *recover'd* some vigour again, but that was of a short duration. (*Philosophical Transactions* Vol. 2, 1667: 563)
- (8) [...] then must we vse such Medicines as we can come by, and the *patient* *recouer* such health as he may. (Bright, *Sufficiencie of English medicines*, 1580: 12)

In (8), the use of *we* with a modal verb of obligation (*must*) implies a persuasive function, which is similar to (4).

When the collocate *sweat* is analyzed, the majority of occurrences (17/21) are verbs (as in (9)).

- (9) Take Sorrell-water fiue or sixe spoonfulls, Treacle-water one spoonefull, London-Treacle one dramme and a halfe, mix them well, giue it warme, and so lay the *Patient* to *sweat*.

(Royal College of Physicians, *Certain necessary directions*, 1636: f. D4r)

Direct imperative formulae, as in (9), can be found in recipes, which are plentiful not only in the category of RECIPE COLLECTIONS in EMEMT but in other categories as well.

The majority of instances (11/12) of the collocate *ease* are used as a noun (as in (10)).

- (10) [...] for the reducing of such a member, hasten to your Bone-setter, but in his absence if such a thing happen in the fingers, toes, or the elbow, two men grasping the member very fast, a third man directing them to pull in the joynt, and the said third man with his right thumbe upon the place, the left hand turning the member into his right place, which the *Patient* will soone finde *ease*, if rightly performed. (Elkes, *Approved medicines*, 1651: 35–36)

In (10), *ease* is used in a general manner with reference to curing the patient or at least alleviating symptoms. This was the most common function in the material but there were others, too, showing more concern for the patient and also mentioning pain: “the *patient* undoubtedly shall finde great *ease*, and void many stones, and much gravel with little pain” (Gray, *Choice manual*, 1653: 4). The use of this collocate indicates that some attention was paid to the patient’s comfort already in the early modern data, especially in connection with painful maladies, such as stones and tooth-ache, but these sentiments cannot be seen as clearly with other collocates in this period.

Many of the instances of the collocate *give* can be found in lexical bundles, such as “*give the patient*” with multiple variations (31 instances), for example, “*give the patient to drink*” (6), “*give the patient thereof to drink*” (4), “and *give it the patient to drink*” (4) and “*give to the patient*” (4). These phrases illustrate the active role of the physician and the passive role of the patient as somebody receiving treatment. Medicinal drinks, posset, etc., were widely administered to patients, more so than other medicines, such as “Fever Pills” and “powder”, which explains the frequency of the word *drink* in the material.

When the plural node *patients* is analyzed in the EMEMT corpus data (see Table 1), similar results to the singular node emerge, at least initially. For example, the word *strength* is the strongest collocate for both the singular and plural form and the contexts of use are the same for the singular and plural node dealing with proper medications and treatments. However, two authors are responsible for 8/10 instances, which skews the results.

The contexts for the collocate *drink* have to do with administering medicinal drinks to patients or referring to what patients are allowed to drink (as in (11)).

- (11) But because some *Patients* cannot *drink* any considerable quantity of water, without nauseating or other inconveniency to their Bodies, the Physician may accommodate the quantity of water to his own liking, provided he order but enough to dissolve it well, that it may not taste too strong.

(Packer, *Mineralogia*, 1693: 7)

Collocate extraction for the plural node *patients* also produces personal pronouns. The possessive collocate *their* is a curious case as most of the examples (21/33) are not about patients and, for instance, their afflictions, but concern other physicians and “their patients”. These examples can be very negative referring to the ill practices of unskilled physicians (as in (12) and (13)).

- (12) How is it possible for these Phisitions, to do any lesse then kyll *theyr pacientes*, but it maketh y^e lesse matter, seinge they be hyred to it, wyth golde and syluer.
(Langton, *Introduction into phisycke*, 1545: 10v)

- (13) But if thei bee menne of knowlege, and thus abuse *their pacientes*, thei are worthy to be punished, as malicious murtherers: [...]

(Bullein, *Bulleins bulwarke*, 1562: 8v)

The ongoing changes in the field of medicine resulted in these outbursts of negativity against other physicians. The early modern period can be characterized as a period of constant disputes and controversies over medical theories and methods of treatment, such as Paracelsian remedies, Helmontianism etc., which is reflected in the material (see Mikkeli & Marttila 2010: 26–27).⁶

The collocate *medicines* does not have the node word *patients* in its immediate proximity, see (14).

- (14) [...] when other cryed up *Medicines* have failed both *Patients* and Physitians it's a Medicine not only of general use (in all families) to be in a rediness, and especially to those who live near the seas, who are so much subject to the Scurvy, and agues.

(Nendick, *Popular pill*, 1674: 12–13)

Nevertheless, the connection between the collocation and the node is clear in all examples in the data.

The collocate *our* occurs especially in the form of *our patients*, but a majority of instances are from one text, the other three texts having one occurrence each. Overall, the first-person plural pronoun is preferred by this author, as in (15).

- (15) But to enquire into the first of these, which advises us to let *our Patients* blood according to their strength, 'tis certain that must never be exceeded; yet there are a great many designs in bleeding, in which that is not so often to be respected.

(Cockburn, *Continuation of the account of distempers*, 1697: 7)

Here the phrase *our patients* is more restricted referring to other physicians (see “specific exophoric *we*” in Wales 1996: 58–63). The plural possessive form “patients” is not indicated in writing during this period, but it is evident from the context.

The results are similarly uneven for the collocate *my*, which seems to be favored by some authors, in this case Colbatch, who is represented by two texts in the corpus from the same year (see (16)).

- (16) By this Method I have brought *my Patients* thro the Distemper, without scarce so much as any uneasiness, or being confin'd to their Beds, or the least fear of a secondary Fever, which frequently proves most fatal.

(Colbatch, *Physicomedical essay concerning alkaly and acid*, 1698: 114)

6. The late modern period also witnessed controversies: inoculation against the smallpox was an important theme as was the theoretical debate over the origin and development of new life in the womb (Pahta 2019: 272). Polemical topics that were of interest to the wider public were discussed to a growing extent in print media, especially the *Gentleman's Magazine* (Taavitsainen 2019: 144).

The author uses the phrase “my patients” in connection with descriptions of successful treatments whereas the contexts of use for the other three authors are neutral.

The most frequent examples of collocate *have* (22 instances) occur as auxiliaries, as in (17).

- (17) For *Patients have* been reduced to that desperate Condition in the Opinion of By-standers, as no appearance of life remained, and yet by God’s blessing directing to the right means, they have beyond hopes recovered, so that none ought to be forsaken before they are dead;

(Chamberlen, *Practice of physick*, 1694: 46–47)

This general advice in (17) was given to persuade readers – consisting of other physicians in the case of this learned text – about the usefulness of any remedy instead of none.

Another common function, when *have* is used as a main verb, indicates what the patients are afflicted with, for instance, “when their Pacients haue the Scuruey” (Wateson, *Cvres of the diseased*, 1598: C3v). The author, that is, the physician, also surfaces with the collocate *have*, for instance in (18).

- (18) And in the Small-Pox that has raged during this present Year 98, I *have* not lost above three *Patients*, and the occasion of their Deaths, as I have all the reason in the World to believe, proceeded from Irregularities committed by their Friends before I was called in, [...].

(Colbatch, *Physicomedical essay concerning alkaly and acid*, 1698: 108)

Of the nine instances of the collocate *such*, six occur as “such patients”, which is a type of reference that appears to be idiolectal as two authors use it twice. These references are neutral, for instance, “... invented merely for the benefit of *such Patients* as cannot endure to swallow liquid drinks” (Charas, *Royal pharmacopoea*, 1678: 60).

The collocate *let* is polysemous as two occurrences deal with letting blood, which skews the results a little. The main function is imperative, as in (19).

- (19) Of any of these Juleps I *let* my *Patients* drink as oft and as freely as they will; drinking likewise small Beer with Juice of Oranges in it, in as large quantities as they please: [...].

(Colbatch, *Physicomedical essay concerning alkaly and acid*, 1698: 110)

This description in (19) is of the kind that typically follows recipes with detailed instructions on the use of the medicine.

The collocate *his* occurs only in phrases “his patients”, often in case narratives or in definitions about the qualities of a good a physician or a surgeon, commonly found in early modern medical texts (as in (20)).

- (20) And ought to eschew al desperate cures, and ought to be gentle to *his patientes*, well willyng to is company, wittie in prognosticating, chast, sober, meeke and mercifull, not auaricious, nor extorcionier for money, but after the capacitie of the Patient, and the facultie receiued, competent and moderate reward.

(Chauliac, *Gvydos qvestions*, 1579: 4v)

Guy de Chauliac was a very influential medical authority from the 14th century, whose works continued to be published well into the early modern period.

4.2 The Corpus of Late Modern English Medical Texts (1700–1800)

The four strongest collocates for the singular *patient* in the LMEMT corpus – *complains*, *aversion*, *alarming*, and *weakened* – have negative denotations that require further scrutiny (see Table 2).

The instances of the collocate *complains* show little variation as they are all of the type “the patient complains of ...”, that is, express dissatisfaction because of symptoms or even pain (as in (21)).

- (21) The *patient complains* of a more than usual degree of heat, a pain and oppression of the breast, especially after motion; [...].

(Cole, *Ladys complete Guide*, 1788: 521)

Example (21) is from the category of health guides, which is a popular genre dating back to the Middle English remedy texts.

Table 2. The strongest collocates for the singular node *patient* (on the left) and the plural *patients* (on the right) in LMEMT

LMEMT							
Collocates	Coll. freq.	Texts	MI	Collocates	Coll. freq.	Texts	MI
<i>Patient</i> (1,813)				<i>Patients</i> (601)			
<i>complains</i> (22)	9	6	8.875	<i>labouring</i> (64)	6	5	8.343
<i>aversion</i> (31)	6	5	7.796	<i>admission</i> (80)	5	3	7.758
<i>alarming</i> (49)	6	6	7.135	<i>visit</i> (91)	5	4	7.572
<i>weakened</i> (45)	5	5	6.995	<i>fifty</i> (104)	5	5	7.379
<i>finds</i> (72)	8	7	6.995	<i>admitted</i> (149)	7	4	7.346
<i>saved</i> (48)	5	5	6.902	<i>hospitals</i> (149)	6	4	7.124
<i>feels</i> (51)	5	5	6.814	<i>home</i> (174)	6	4	6.900
<i>visited</i> (52)	5	5	6.786	<i>house</i> (334)	11	4	6.833
<i>blooded</i> (57)	5	5	6.654	<i>attend</i> (222)	7	7	6.771
<i>quiet</i> (76)	6	5	6.502	<i>hospital</i> (365)	11	5	6.705

The contexts of the collocate *aversion* are similar to the collocate *complains*, dealing with the patient's dislike of medication or other treatment, with the subject position in the clause giving more emphasis to the patient (as in (22)).

- (22) If the medicine be in itself unpleasant, if the *patient* has an *aversion* or dislike to the taking of it, or if the application or taking be difficult and troublesome; these call for the exhibition of as few doses as possible.

(Gaubius, *Complete Extemporaneous Dispensatory*, 1741: 31)

Two examples are from Berkenhout's treatise on rabies where discussing aversion towards water is to be expected as it was considered one of the symptoms of the disease, although the author challenges this belief: "It rarely happens that the *patient* has any *aversion* to water or other liquid" (*An essay on the bite of a mad dog*, 1783: 5).

The collocate *alarming* refers to either an alarming case or situation or something alarming the patient (as in (23)).

- (23) On the other hand, if the bleeding should happen to return, though on account of the size of the vessels it might not be dangerous, yet even here more blood might be discharged than it would be prudent to lose; a fresh hæmorrhage is always *alarming* to the *patient*; and as momentary pain seems to be the only inconvenience, perhaps many surgeons will think it best to leave their patients in perfect security: [...]

(Kirkland, *Suppressing hæmorrhages*, 1763: 54)

Here the concerns of the patient are clearly taken into account by the author-physician.

The collocate *weakened* (see (24)) occurs as a regular verb or as a past participle, associated with a patient's deteriorating condition. The patient is the passive object, but the denotation of the verb arouses some empathy.

- (24) The tobacco-smoke, and the composition containing the infusion of tobacco, occasioned considerable griping, with narcotic effects, and also greatly *weakened* our *patient*, but without producing any evacuation from his bowels.

(*Edinburgh Medical Journal* 4, Vol. 1, 1796: 315)

Analyzing the contexts of these strong collocations for *patient* extracted from the LMEMENT corpus illustrates a shift in focus away from the early modern emphasis of describing various cures moving towards the patient, the patient's emotions and the present state of health.

Most of the occurrences of the collocate verb *finds* describe the patient's health (as in (25)).

- (25) [...] tho' the *patient finds* himself never the better, but perhaps still grows worse: If one specific succeed not, another is tried, and, if that fails, a third, and so on [...] as long as the poor infatuated believing sufferer has any money, strength, or patience left. (Morland, *Causes of chronic diseases*, 1774: 11)

In (25), the role of the patient is stressed as the subject of the clause, and the reference to “the poor infatuated believing sufferer” appeals to the reader. Sometimes the collocate *finds* also refers to the physician: “The Doctor, according as he *finds* the *Patient's* Case to be, sometimes orders a little Blood to be taken away” (Anonymous, *Complete family-piece*, 1736: 29). Overall, the contexts for the collocate *finds* are quite neutral when compared to the other collocates for *patient* in LMEMT.

With the collocate *saved*, the focus is again on the patient or the patient's life as the subject, although the physician is still actively present in the context of situation. The formulaic phrasing of “might have been saved” still puts more emphasis on the patient (see (26)).

- (26) Some may imagine that this Accident was not owing entirely to the Emetic, but a Plethora; and had bleeding preceded, in order to have emptied the Vessels, the *Patient* might have been *saved*: [...]. (Catherwood, *Curing the apoplexy*, 1715: 30)

Sometimes, however, the role as an object of action is reserved to the patient: “taken in such an extraordinary quantity as to bring on immediate death, has frequently *saved the patient*” (*Gentleman's Magazine* Vol. 25, 1755: 29).

Of the five instances of collocate *feels*, all from different authors, four link directly to what the patient feels, especially in terms of experiencing pain or “uneasiness”, as in (27), (28) and (29).

- (27) If the ligature recommended is used, the parts will be less irritated; and whether the absolute security against a future hæmorrhage does not fully compensate for the uneasiness the *patient feels* on this account, may be gathered from what follows. (Kirkland, *Suppressing hæmorrhages*, 1763: 25–26)
- (28) In about six or seven Days after the Inoculation, the *Patient* sometimes *feels* a Kind of a slight Pain in the Inoculated Arm, in his Back and Head, but all those Symptoms soon go off. (Anonymous, *Inoculation made easy*, 1766: 8)
- (29) The *patient* at one time *feels* as if he were burnt with devouring fire, and soon after is sensible of cold, similar to what arises from the application of water. These symptoms are accompanied with loss of appetite, nausea and vomiting, which are soon succeeded by an inexpressible pain. The patient, without shedding tears, utters lamentable cries, and the affected members are either contracted with extreme violence, or remain stretched out with inflexible rigidity. (*Edinburgh Medical Journal* 3, Vol. 9, 1783–1784: 80)

The vivid description of pain in (29) further underscores the focus on the patient. Many of the collocates for the node *patient* in the LMEMT corpus have negative semantic prosodies that can be related to either a noun or a verb.

In comparison, all the instances of the verb *visited* were found in neutral contexts, for instance, “I have *visited* your *patient* the General thrice” (Rollo, *Diabetes mellitus*, 1797: 15). This particular text consists of letters sent to the author by different physicians concerning one patient, which explains the use of the second person pronoun.

The collocate *blooded* was used with imperative verb forms, for instance “let the *patient* be *blooded*” (Taylor, *Family companion*, 1795: 152), which are common in texts consisting of a collection of recipes and various cures. Modal verbs were also utilized in these texts to instruct the readers, for instance “the *patient* must also be *bloodied*” (Theobald, *Every man his own physician*, 1764: 13) or “the *patient* should not be *blooded* till after a month’s purging” (*Gentleman’s Magazine* Vol. 37, 1767: 595). These examples refer to the traditional method of letting blood in order to relieve the patient from superfluous humours or imbalances in the blood. The patient is reduced to an object of a medical procedure performed by a physician. The function here resembles the way in which *purged* was used in the EMEMT corpus.

Of the instances of the collocate *quiet*, most refer directly to the patient (as in (30)).

- (30) To restrain this flux, the *patient* should be kept *quiet* and easy both in body and mind. (Buchan, *Domestic medicine*, 1772: 654)

Advice is also given to a fellow colleague, when preparations are made for the patient to undergo an operation: “give him a small quantity of wine, and you will find that the *patient* will be *quiet*” (Bischoff, *Extraction of the cataract*, 1793: 45). Here the patient is described as someone to be kept quiet so that the physician can practice his work without disturbance.

When the plural node *patients* is examined in the LMEMT corpus, the collocate *labouring* emerges as the strongest collocation (see (31)).

- (31) I have had the management of a great number of *patients labouring* under this obstinate complaint, and have been successful in curing every individual that has been under my care. (Jones, *Nervous fevers*, 1789: 16)

Labouring can be seen as a euphemism for the suffering under various afflictions: all the examples contain a lexical bundle “patients, labouring under ...”, for instance, “under fever” (Currie, *Effects of water*, 1797: 9) and “kinkcough” (*Edinburgh Medical Journal* 3, Vol 1, 1773: 391).

The collocate *admission* can only be found in three texts all of which have a clearly institutional focus (as in (32)).

- (32) Whether they be contagious, or subject in a peculiar degree to corrupt the air and generate pestilential diseases – the danger of their *admission* to the other *patients* is obvious. (Aikin, *Thoughts on hospitals*, 1771: 22)

The use the word *admission* in connection with pestilential diseases might affect the otherwise neutral meaning of the collocate.

The five instances of the collocate *visit* are all located before the object *patients*: for instance, “to *visit patients* in the city of London” (Anonymous, *Constitution of physick*, 1768: 9) and all but one occurrence represent institutional texts in the corpus.

It is noteworthy that even the numeral *fifty* emerges as a collocation to the word *patients* (the exact reference of one of these is *fifty-four*). Examples were distributed quite evenly in the corpus, suggesting that it is not only happenstance, for instance: “the Subscribers to the Dispensaries have *fifty* or an hundred gratis *Patients*” (Anonymous, *Usefulness of the dispensaries*, 1702: 11) and “received into the hospital more than *fifty patients* afflicted d’une gangrene” (*Philosophical Transactions* Vol. 52, 1761: 532). Even if the context in all examples is institutional, connected to hospitals, references to numbers of patients also surface in other kinds of texts, underlining the developments in medicine in the 1700s.

The contexts of the collocate *admitted* have to do with public health. For instance, a text on mental institutions is represented (see (33)).

- (33) Resolved, That it would be convenient to the Public, as well as to the Asylum, if *patients* in affluent circumstances were *admitted*. (Anonymous, *Letter to York lunatic asylum*, 1788: 23)

The collocate also appears in a text on smallpox patients, for instance: “they have lately *admitted* small-pox *patients*” (Lettsom, *Improvement of medicine in London*, 1775: 10). These uses are neutral, but one example seems to acquire negative undertones due to the emotive adjective used to describe patients: “those miserable objects from being *admitted* as in-*patients*” (*Gentleman’s Magazine* Vol. 43, 1773: 184), which is the only occurrence outside the PUBLIC HEALTH category. The collocate *hospitals* can only be found in neutral contexts, for instance, “*patients in hospitals*” (Bell, *Surgeons of Edinburgh*, 1800: 9). Not all instances are from institutional texts – two out of six are from texts dedicated to specific diseases.

Half of the instances of the collocate *home* (3/6) are found in Lettsom’s *Improvement of medicine in London* from 1775, where all references are to the compound *home-patients*. The idea of treating patients at home versus in the hospital is present in other instances as well, in texts on specific diseases or cures, for instance, when the services of a physician are announced (as in (34)).

- (34) The Author will attend *patients* at *home* and abroad. (Trinder, *Application of oils*, 1797: 30)

This phrase is located at the end of the text and it has an advertising function.

The collocate *house* can mostly be found in lists, such as “*Patients in the House 127*” (London Hospital, *London infirmary*, 1742: 11) and the “number of *patients* in the *house*” (Bowen, *Bethlem hospital*, 1783: 6). The meaning of *house* in all these examples is, in fact, ‘hospital’ and the context is mostly institutional (in 6/7 instances).

The collocate *attend* is used in two ways: with reference to patients attending or not being able to. It also refers to physicians attending (to) something as in (34). Of the seven occurrences, four have an institutional origin and others deal with cures or diseases.

The collocate *hospital*, in the singular, occurs in generic phrases “*patients* in an *hospital*” or “*hospital patients*”. Most of the examples are from institutional texts (7/11), but not all, for example: “safe and *sufficient* experiments might be made on *patients* in an *hospital*” (*Gentleman’s Magazine* Vol. 37, 1767: 461). This resembles the context in which the collocate *house* was used.

5. Discussion

When collocations for the words *patient* and *patients* are assessed in the early modern and late modern periods, significant differences can be detected. With the older material in the EMENT corpus, similarities in the use of collocates between the singular *patient* and the plural node *patients* are apparent; for example, the collocate *strength* is used in the same fashion with both the singular and the plural node. In the same period, collocates referring to various treatments – *smoke*, *purged*, and *sweat* – occur with the singular *patient* while the plural *patients* is accompanied with hyperonyms *medicines* and *drink*, which indicate that the contexts of use were not very different. In the LMENT corpus, however, similarities between the singular and plural node are few and far between, the words *visit* and *labouring* being the only exceptions, the latter of which is used to create compassion towards patients suffering from an illness; this function can be found with a number of the collocates for the singular node. All the other nine collocates for the plural *patients* relate to hospital life, from *admission* to *hospitals* – even the numeral *fifty* emerges as a collocate to refer to the number of patients. Therefore, it is tempting to suggest that in the late modern period both the singular and the plural node *patient/s* had developed specific semantic prosodies that were previously not present.

Considerations about the comparability of the two corpora need to be raised here. The register of medical writing underwent changes in time – genres evolved and entirely new sub-genres, such as the medical journal, developed and new topics had to be addressed – is it then possible to draw comparisons between the two datasets? The changing genre conventions and the birth of new genres and topics unquestionably had an effect on the language, which is why corpus categories were taken into

account in the analysis. A case in point is the category of institutional texts, which is overly represented in the collocations for the node *patients* in the late modern period. Although some of the collocates occur almost solely in institutional texts (*admission, visit, house*), a number of them can also be found in other kinds of texts (e.g. *fifty, admitted, home, attend, hospital*). These results emphasize the role of institutional texts as a new (sub)genre, and perhaps also underline some of the differences in the constellation of categories of the EMENT and LMENT corpora. As medical writing evolves and specializes, producing new genres, this is reflected in medical discourse and the compilation of corpora should, in turn, mirror these developments.

A closer examination of the results shows that some of the collocates point towards idiosyncrasies of certain authors and texts and to their slight overrepresentation in the results, which was nevertheless moderated by the fact that collocates present in less than three different texts were excluded. This problem is more pronounced in the EMENT corpus and specifically with the singular node *patient* and the collocates *strength* and *let* as around third of all instances for these collocates were from one author. Collocates can be informative of prevalent ideas and patterns of language use, but the findings need to be carefully interpreted if generalizations are to be made. In the case of late modern data, it can be speculated that when the general environment became more susceptible towards expressing compassion towards patients, some authors were more prone towards using these expressions. The quality of collocates as preferred phrases and word-choices by an individual author is not an exhaustive explanation. The genre is also an important factor in the choice of vocabulary as well as the topic of the text. It is relevant to discuss a patient's feelings in a text on painful afflictions or for instance, on hemorrhages, which can be frightening for the patient – this became apparent in the analysis of the data. The collocates used in emotive contexts in late modern data, especially with the singular node *patient* and with the collocates *complains, aversion, alarming, weakened, and feels*, all convey negative semantic prosodies in the material. In contrast, in early modern texts the patient is described as the object of action, commonly with imperative verb forms. When during this earlier period the patient acts as the subject in the clause, it had to do with recovering (*recovered/ recover*) from illness with the singular node *patient* or with suffering from diseases with the plural *patients* (with the collocate *have*).

6. Conclusion

This chapter set out to evaluate, with the help of collocation analysis, how the attitudes of physicians towards their patients change in a long diachrony (1500–1800). The quantitative method provided a wealth of material, which needed to be qualitatively assessed making the method quite time-consuming. The analysis suggests

that collocations can reveal changes in attitude within society, which were then reflected in the field of medicine and in medical writing. Many of the ideas emerging from the material would have been difficult to notice otherwise, especially the larger shift during the 18th century from the patient as an object receiving treatments to *experiencer*. References to medicines and cures were prevalent in the early modern data, but lost some of their significance in the latter period. In turn, the collocates for the plural node *patients* in the LMEMT corpus show the new importance of hospitals and the advances in public health during the 18th century. At the same time, the negative connotations of some of the collocates, especially with the singular node, emphasize the patients' experiences as well as the generally agreed ideals of polite society reflecting the ideas of Enlightenment: the urge to treat patients more humanely.

References

Corpora and software

- AntConc = Anthony, Laurence. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo: Waseda University. <<http://www.laurenceanthony.net/software>>
- EMEMT = *Early Modern English Medical Texts: Corpus* [CD-rom]. 2010. Compiled by Irma Taavitsainen, Päivi Pahta, Martti Mäkinen, Turo Hiltunen, Ville Marttila, Maura Ratia, Carla & Jukka Tyrkkö. Amsterdam: John Benjamins.
- LMEMT = *Late Modern English Medical Texts. Corpus* [CD-rom]. 2019. Compiled by Irma Taavitsainen, Turo Hiltunen, Anu Lehto, Ville Marttila, Päivi Pahta, Maura Ratia, Carla Suhr & Jukka Tyrkkö. Amsterdam: John Benjamins.
- VARD = Variant Detector, version 2.3.

Other references

- Atkinson, Dwight. 1992. The evolution of medical research writing from 1735 to 1985: The case of the *Edinburgh Medical Journal*. *Applied Linguistics* 13(4): 337–374.
<https://doi.org/10.1093/applin/13.4.337>
- Barnbrook, Geoff, Mason, Oliver & Krishnamurthy, Ramesh. 2013. Collocation and language theory: Recent developments. In *Collocation: Applications and Implications*, 147–173. London: Palgrave Macmillan. https://doi.org/10.1057/9781137297242_7
- Baron, Alistair & Rayson, Paul. 2009. Automatic standardization of texts containing spelling variation: How much training data do you need? In *Proceedings of the Corpus Linguistics Conference, CL2009, University of Liverpool, UK, 20–23 July 2009*, Michaela Mahlberg, Victorina González-Díaz & Catherine Smith (eds). Liverpool: University of Liverpool Press. <<https://eprints.lancs.ac.uk/id/eprint/42529/>> (7 November 2019).
- Biber, Douglas & Conrad, Susan. 1999. Lexical bundles in conversation and academic prose. In *Out of Corpora: Studies in Honour of Stig Johansson*, Hilde Hasselgård & Signe Oksefjell (eds), 181–190. Amsterdam: Atlanta.

- Brezina, Vaclav, McEnery, Tony & Wattam, Stephen. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2): 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>
- Brown, Michael. 2011. Polite and ornamental knowledge: Medicine and the world of letters. In *Performing Medicine: Medical Culture and Identity in Provincial England, c. 1760–1850*, 48–81. Manchester: Manchester University Press. <https://doi.org/10.2307/j.ctt21216ph.8>
- Church, Kenneth W. & Hanks, Patrick. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22–29.
- Durrant, Philip & Doherty, Alice. 2010. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory* 6(2): 125–155. <https://doi.org/10.1515/cllt.2010.006>
- Evert, Stefan. 2005. The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD dissertation, University of Stuttgart. <https://doi.org/10.18419/opus-2556>
- Gries, Stefan T. 2013. 50-something years of work on collocations: What is or should be next.... *International Journal of Corpus Linguistics* 18(1): 137–165. <https://doi.org/10.1075/ijcl.18.1.09gri>
- Gross, Alan G., Harmon, Joseph E. & Reidy, Michael S. 2002. *Communicating Science: The Scientific Article from the 17th Century to the Present*. Oxford: OUP.
- Hiltunen, Turo. 2010. Category 6. Philosophical Transactions. In *Early Modern English Medical Texts: Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 127–131. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.160.04cat6>
- Kilgariff, Adam. 1996. Which words are particularly characteristic of a text? A survey of statistical approaches. In *Language Engineering for Document Analysis and Recognition*, Brighton, England, April. AISB Workshop Series, 33–40.
- Klein, Lawrence. 2002. Politeness and the interpretation of the British eighteenth century. *The Historical Journal* 45(4): 869–898. <https://doi.org/10.1017/S0018246X02002716>
- Lehto, Anu, Baron, Alistair, Ratia, Maura & Rayson, Paul. 2010. Improving the precision of corpus methods: The standardized version of *Early Modern English Medical Texts*. In *Early Modern English Medical Texts: Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 279–290. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.160.11leh>
- Lehto, Anu. 2019. Public health. In *Late Modern English Medical Texts: Writing Medicine in the Eighteenth Century*, Irma Taavitsainen & Turo Hiltunen (eds), 307–315. Amsterdam: John Benjamins.
- Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and Technology: In Honour of John Sinclair*, Mona Baker, Gill Francis, John Sinclair & Elena Tognini-Bonelli (eds), 157–176. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.64.11lou>
- McEnery, Tony & Hardie, Andrew. 2013. The history of corpus linguistics. In *The Oxford Handbook of the History of Linguistics*, Keith Allan (ed.), 727–746. Oxford: OUP. <https://doi.org/10.1093/oxfordhb/9780199585847.013.0034>
- Mikkeli, Heikki & Marttila, Ville. 2010. Change and continuity in early modern medicine (1500–1700). In *Early Modern English Medical Texts: Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 13–27. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.160.02mik>
- Nevalainen, Terttu & Tissari, Heli. 2010. Contextualising eighteenth-century politeness: Social distinction and metaphorical levelling. In *Eighteenth-century English: Ideology and Change*, Raymond Hickey (ed.), 133–158. Cambridge: CUP. <https://doi.org/10.1017/cb09780511781643.009>

- Nilsson Björkenstam, Kristina, Gustafson Capková, Sofia & Wirén, Mats. 2014. The Stockholm University Strindberg Corpus: Content and possibilities. In *Strindberg on International Stages/ Strindberg in Translation*, Roland Lysell (ed.), 21–40. Newcastle upon Tyne: Cambridge Scholars.
- OED Online. 2018. <<http://www.oed.com>> (December 27, 2018). Oxford: OUP.
- Pahta, Päivi. 2019. Midwifery. In *Late Modern English Medical Texts: Writing Medicine in the Eighteenth Century*, Irma Taavitsainen & Turo Hiltunen (eds), 271–278. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.221.11.2d.pah>
- Pahta, Päivi & Ratia, Maura. 2010. Category 2. Treatises on specific topics. In *Early Modern English Medical Texts: Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 73–99. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.160.04cat2>
- Pahta, Päivi & Taavitsainen, Irma. 2010. Introducing *Early Modern English Medical Texts*. In *Early Modern English Medical Texts: Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 1–7. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.160.00toc>
- Pollach, Irene. 2012. Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods* 15(2): 263–287. <https://doi.org/10.1177/1094428111417451>
- Porter, Roy. 1985. Lay medical knowledge in the eighteenth century: The evidence of the *Gentleman's Magazine*. *Medical History* 29: 138–168. <https://doi.org/10.1017/S0025727300043970>
- Porter, Roy. 1986. Laymen, doctors and medical knowledge in the eighteenth century: The evidence of the *Gentleman's Magazine*. In *Patients and Practitioners: Lay Perceptions of Medicine in Pre-industrial Society*, Roy Porter (ed.), 283–314. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511563690.010>
- Porter, Roy. 1991. Cleaning up the great wen: Public health in eighteenth-century London. *Medical History Supplement* 11: 61–75. <https://doi.org/10.1017/S0025727300071118>
- Ratia, Maura. 2019. Diseases. In *Late Modern English Medical Texts: Writing Medicine in the Eighteenth Century*, Irma Taavitsainen & Turo Hiltunen (eds), 243–249. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.221.11.2a.rat>
- Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1): 23–55. <https://doi.org/10.1075/fol.2.1.03stu>
- Taavitsainen, Irma. 2011. Medical case reports and scientific thought-styles. *Revista de Linguas para Fines Especificos* 17: 75–98.
- Taavitsainen, Irma. 2019. General periodical: *The Gentleman's Magazine*. In *Late Modern English Medical Texts: Writing Medicine in the Eighteenth Century*, Irma Taavitsainen & Turo Hiltunen (eds), 327–336. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.221.11.8.taa>
- Taavitsainen, Irma & Tyrkkö, Jukka. 2010. Category 1. General treatises and textbooks. In *Early Modern English Medical Texts: Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 65–72. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.160.04cat>
- Taavitsainen, Irma, Hiltunen, Turo, Lehto, Anu, Marttila, Ville, Ratia, Maura, Pahta, Päivi & Tyrkkö, Jukka. 2014. *Late Modern English Medical Texts 1700–1800: A corpus for analysing eighteenth-century medical English*. *ICAME Journal* 38(1): 137–153. <https://doi.org/10.2478/icame-2014-0007>
- Wales, Katie. 1996. *Personal Pronouns in Present-day English*. Cambridge: CUP.

Semantic neology

Challenges in matching corpus-based semantic change to real-world change

Antoinette Renouf

Birmingham City University

This study deals with contemporary real-world social change and its linguistic expression by means of neosemes, with the help of the WebCorpLSE automated collocation analysis software. The study briefly outlines the history and pattern of diffusion of each of the five user-defined neosemes, noting where possible the tipping point of popularization. We examine these modern neosemes in a series of case studies, focussing on the challenges faced in this task, for instance in the form of rarity of occurrence, as emergent or ephemeral formations. The functionality of the software is examined in each case, in order to demonstrate various measures which can be undertaken to allow or enhance its performance in teasing out collocational profiles from sparse data.

1. Introduction

In this chapter, the focus is on the emergence of semantic neologisms (henceforth *neosemes*, coined by Stockwell 2000)¹ in a news corpus in response to changes in society. We define semantic neology (henceforth *neosemy*) as the polysemic addition of a new sense to an existing word form: new in the broadening or narrowing of sense, or by analogy, metaphor, metonymy, among other changes.² This study complements earlier work on the life cycle of words (Renouf 2013), which focussed on lexical neology in news text, but dealt only cursorily with neosemy. It also builds on a recent evaluation of the performance of our automated system of identifying neosemes by collocational change (Renouf 2014).

1. "It is important, firstly, to distinguish between new word-shapes and new meanings attached to existing words; [...] I will refer to the second type as neosemes and will discuss 'neosemy' (to coin a word) first" (Stockwell 2000: 109).

2. Also, synecdoche, hyperbole, litotes, and evaluative connotation.

In the 1990s, we developed the automated systems AVIATOR (Renouf 1993), ACRONYM (Renouf 1996) and APRIL (Renouf 2010), which identify candidates for neosemy according to the criterion of collocational change (in the knowledge that not all polysemic development is reflected collocationally). That is to say, a significant statistical change in the collocational profile of a word is deemed to indicate a new sense. We considered just the linguistic parameters of first-order collocation and patterns of frequency, along with statistical measures of relative and significant frequency. This was an experimental compromise to exploit readily-identifiable surface textual features, thereby streamlining the identification process while disregarding less accessible features such as second-order collocation (Mollet, Wray & Fitzpatrick 2011)³ or collocational networking (Brezina, McEnery & Wattam 2015) which could also play a role. The surface approach has shown itself to be surprisingly successful, though it cannot differentiate between subtleties of meaning change such as use and reference, or stylistic uses such as irony or euphemism.

Interest in the automatic identification of semantic change has begun to emerge in recent years within other research groups (Boussidan & Ploux 2011; Cabré & Nazar 2012; Cartier 2016; Reutenauer 2012). These groups are still grappling with the conceptualization of semantic change criteria, and its semi-automatic operationalization by relevant software, and this on a diachronic corpus.

Against this background, our aim is to manually select words manifesting semantic change which reflects current real-world change, and to assess and test the challenges that such neosemes present to the University of Birmingham Research and Development Unit for English Studies (RDUES) automatic analytical software, in identifying that change in the UK news corpus solely by collocational means.

2. Data and methods

2.1 Data and tools

The data comprise an accumulating flow of *Independent* and *Guardian* newspapers from 1984 to 2017 (ongoing), of around 1.7 billion words of news text. These data are fed in monthly time chunks through software filters, which compare them to identify ongoing significant collocational change, and are thence fed into an evolving collocate bank. These tools were developed within the 1990–1993 AVIATOR

3. According to Mollet et al. (2011: 89), “capturing second-order patterns is computationally extremely greedy.”

project and the 1997–2000 APRIL project, and are located within the 2006-onward WebCorp Linguist's Search Engine (WebCorpLSE) toolset.

WebCorpLSE tools (Kehoe & Gee 2009) provide visual displays of analyzed data. We have worked with a series of graphic representations of linguistic reality over the years. For neosemy, we currently favour two display tools. One creates a 'time series graph' (henceforth *time graph*), which plots the changing corpus frequency of a word across time on a graph. The graph is scalar for time and frequency, and the dotted plotting line can be presented as simple frequency per million words, with a solid smoothing line for ease of interpretation. This reveals a word's life cycle (Renouf 2007, 2013), and allows the testing of hypotheses about increased frequency and topic salience, productivity and semantic change. The second graphic tool is a *heat map*, which represents the changing collocational significance scores across time for each collocate of a node word, by blocks of colour of varying intensity. This tool shows the evolution of the collocational profile of a word across the years, which is thus indicative of the path of emergence of a new sense.

The system stores information capable of suggesting lexical candidates for newness, productivity, collocational, and semantic change, and so on, but it is not pro-active in terms of making suggestions to the user. The corpus and software analytical systems currently contain 1.7 billion running words and 4.4 million word types. The user decides which of these words to analyze on the basis of existing knowledge or hypothesis, and the system assembles the relevant range of information in store for each selected type. The system is currently under development, in the first instance to become pro-active in identifying word frequency change types.

2.2 Tracking the neosemes

In the selection of neosemes, previously studied items were disregarded, while new drivers for recent social change were considered. A range of possible candidates was found by an extensive reading of US and UK news and tested against the corpus lexicon, a fairly time-consuming task, with many candidates being rejected. The list was narrowed down to the following: *birther*, *normalis/zation*, *cougar*, *snowflake*, and *ghosting*. Between them, these neosemes cover the following areas of change: political theory and its proponents (*birther*); a socio-political and mental/psychological process (*normalis/zation*); recent social stereotypical issues (*cougar* and *snowflake*); and a technologically newly-enabled social process (*ghosting*).

Each case study will briefly sketch the provenance of the target neoseme, both its social background and the relevant etymology. We then consider the linguistic circumstances of each neoseme, estimating the probable challenges it presents to the automated collocational change software. The data will be calibrated for best

results. For example, the retention of upper/lower case, singularization/pluralization and lemmatization will depend on the commonality of sense between the individual forms of each neoseme.

The next stage is to trace the evolution of the originating social changes prior to and during the life of each selected neoseme, to establish provenance and pattern of diffusion via the media, social networks, and so on. In defining and tracking senses, the *Oxford English Dictionary* has been the first port of call; followed by major online dictionaries, as well as *Urban Dictionary* and *Wiktionary* for recent senses. Factual information has been drawn from online websites including *Wikipedia*, and physical reference sources. The etymology provided for each neoseme takes into account previous senses which impinge collocationally upon the target neoseme: that is, the sense of the base form, rather than root morpheme (Bauer 1983: 21).

Attention is then paid to the probable challenges presented by the selected neosemes. They all engender only sparse data as new formations which do not achieve full maturity before the corpus ends in December 2017, or as rare items with ephemeral reference. Associated effects on their performance include the boundedness of the denoted event, the degree of polysemy surrounding the neoseme, and the consequent likely ease or difficulty in its collocational recognition.

Before processing, adjustments are made to search parameters, to enable or optimize the process. For greater recall, these include extending the collocational span to a maximum of ± 4 , combining or lemmatizing the derived and inflected forms of the neoseme; and allowing upper and lower-case typographical variants. For increased precision, they include restricting the collocational span; restricting search to corpus time chunks recording peak use; tailoring the selection of derived and inflected forms of a neoseme; using Z-score statistics to promote or demote collocational profiles; focussing the search string with lexical or syntactic pattern specification; and as a last resort, tailoring the search by manually weighting the search string with required collocates.

A search string for each neoseme is then submitted to the WebCorpLSE automated collocation change analysis software, which produces quantified output and presents it graphically to the user as evidence of how far it succeeds in reflecting the sense change.

3. Case studies

3.1 Case study 1: *Birther*

The neoseme *birther* and its derived forms, including *birtherism*, denote a political theory and its proponents. Specifically, *birtherism* is an example of social issues which have been engendered by the political turmoil surrounding doubts about the country of birth of Barack Obama, and thus his eligibility to campaign for re-election from 2011.

3.1.1 *Provenance*

There is no OED entry for *birther/birthers*. In Wiktionary,⁴ we find two senses, the first of which is as follows:

Sense 1: 1933. One who gives birth. (A noun derived from *birth*.) The term *birther* emerged in the 1930s, to denote the practitioners of a UK doctrine militating for non-hospitalized and non-medicalized methods of childbirth.

A manual search of the lower-case noun forms *birther* and *birthers* shows that, out of a total 506 corpus occurrences, they convey only a total of 44 occurrences of Sense 1 (33 of which in combination with compounding elements (*natural* 10, *home* 6, *free* 6, *child* 3, *hi-tech* 3, and *lotus* 2, *serial* 1, *baby* 1, *pro* 1).⁵

The first and final occurrences in our corpus for *birther/s* in Sense 1 are:

- (1) the cut-hungry Health Minister will take the natural *birthers* at their word (05/1985)
- (2) The “freebirthers” in your article... live in cloud cuckoo land. (09/2013)

and only four instances⁶ of *home#birthers* appear after 2008.

For Sense 2 of *birther*, Wiktionary offers:

Sense 2: 2008. A believer in the conspiracy theory that Barack Obama was not born in the US, and therefore ineligible for presidency due to national law.

Conspiracy theories regarding Obama’s presidential ineligibility were propagated for years leading up to, during and after his campaign and eventual election. In 2004, Andy Martin said he was a “crypto-Muslim”. In 2008, Clinton supporters

4. <<https://en.wiktionary.org/wiki/birther>>

5. Disregarding *rebirther/s* 11, as a different semantic case.

6. In fact, only two instances, of two pairs of erroneous printing duplications.

questioned his citizenship (see Rutenberg 2008). Even in 2011, when a partial birth certificate was produced stating that Obama was born in Hawaii, Donald Trump reportedly reignited the issue by saying that he still had doubts about Mr Obama’s account; and there was continuing pressure from “fringe theorists” or “Birthers”. Hawaii then produced a full birth certificate. Despite the issue having been resolved, Donald Trump reportedly continued to express doubts between 2011 and 2016 (see Barbaro 2016).

The neoseme *birtherism* began to appear in the US media from March 2008, but is not found in our UK corpus until July 2009, as in (3).

- (3) the nutters in the Obama “birther” movement. These are the people who swear Obama wasn’t born in America (07/2009)

The derived noun forms *B/birther*, *B/birthers*, *B/birtherism*, *Birtherish*, *birther-ish* convey this sense in the corpus. The delay in its reaching UK news text was possibly due to UK journalistic pre-occupations with the 2008 global crash, a volcanic eruption, and UK-specific issues such as *flipping* and *downsizing*.

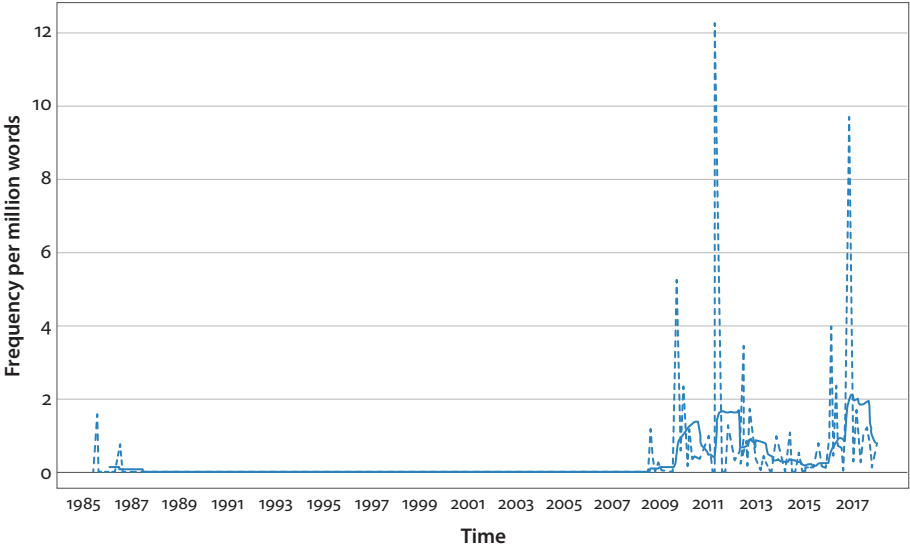


Figure 1. Time graph of 667 occurrences of Sense 2 [*Birther/birther/Birthers/birthers/Birtherism/birtherism/birtherish/Birther-ish*]

A time graph estimates the corpus life cycle for the neoseme *birtherism* in Figure 1. The search string itemized in Figure 1 produces approximately 667 occurrences, of which approximately 44 are Sense 1. Thus, the time graph in Figure 1 is only approximate, but it does indicate low peaks of activity, around 1.5–2 per million

words, from 2009 to 2011, at the official resolution of the Obama birth issue, with a continuing low presence thereafter, until a new growth in use, from late 2016 until the end of 2017, when evaluation of the past event takes place.

Our UK corpus news text references remain essentially only reports of the US affair and its ramifications, indicating that the neoseme has not been assimilated into British cultural reference or use.

3.1.2 *Assessment of challenge*

Sense 1 of *birther* (derived forms *birther/s*) is rare in the corpus, as said, occurring only 44 times. It denotes a relatively time-bound UK event, ceasing to feature in the corpus after 2008 (except for two references), leaving a clear gap for the Sense 2 neoseme *birtherism* to emerge from 2009. This runs until the end of the corpus, despite the birtherism issue being legally resolved in 2011. The relevant forms amount to 506 occurrences, in principle furnishing sufficient data to support an automated collocational profile. The prediction is thus that the neoseme will succeed in matching the real-world social change with a corpus-based collocational representation.

3.1.3 *System performance*

For Sense 1, *birther* denoting women undergoing un-medicated childbirth, the 44 occurrences are insufficient to generate an automated collocational profile. The commoner Sense 2 neoseme⁷ has 667 occurrences, suffers no collocational competition from Sense 1, and has sufficient relevant collocates to process, so that the system is able to retrieve a successful collocational profile.

This result is enhanced by focussing on the sequence of data time chunks representing the peak use of the neoseme between 2009 and 2012. In Table 1, we see the result, which suggests the new meaning through the recurrent collocates, including: *movement*, *conspiracy*, *theories*, *Donald*, *controversy*, *Obama*, *Trig* (name of a Birther), *Trump's*, *certificate*, and *Trump*.⁸ A narrow collocational span of ± 2 to the node word is sufficient to make the case; a wider span yields more of the same. Equally, recourse to Z-score statistics (not shown here), which would re-order the output to promote 'surprise' elements, does not improve the result.

7. *Birther/birther/Birthers/birthers/Birtherism/birtherism/birtherish/Birther-ish*

8. A collocational frequency count of as low as three occurrences, for *certificate* and *Trump*, is not ideal, but the data are sparse, as said from the outset.

Table 1. Sense 2 collocational profile for [*birther*_2009–2012*], collocation span $\pm 2^*$

Word	L2	L1	Total	R1	R2
<i>movement</i>			32	31	1
<i>conspiracy</i>			13	12	1
<i>so-called</i>		12	12		
<i>issue</i>			9	9	
<i>Donald</i>	4		8	1	3
<i>controversy</i>			8	8	
<i>Obama</i>	1	5	7		1
<i>Trig</i>		7	7		
<i>Herman</i>			5		5
<i>plan</i>	5		5		
<i>comments</i>			5	5	
<i>nonsense</i>			5	5	
<i>theories</i>			5	1	4
<i>backing</i>	5		5		
<i>Trump's</i>		5	5		
<i>claim</i>			4	2	2
<i>theory</i>			4	2	2
<i>believe</i>	1		4	1	2
<i>debate</i>			4	4	
<i>certificate</i>	3		3		
<i>Trump</i>	1		3		2

* Responding to a reviewer query, if the search terms for Sense 1 and Sense 2 are combined, the collocational profile at ± 2 to left and right of the node word is very similar to the one in Figure 3, since the 44 Sense 1 occurrences are drowned out or severely suppressed by the 677 occurrences for Sense 2. However, if the collocational span is reduced to ± 1 to left and right, the top collocate from Sense 1, namely *natural*, creeps in at rank 8.

By September 2016, continuing heated political debate, as well as a journalistic desire to keep the story alive by retrospective analysis, ensure a presence of the neoseme until the end of the corpus. This post-2016 trace account of the no-longer topical neoseme can be seen in the collocational profile in Table 2, where it reflects a negative evaluation (see *attack, lie, lies, accusations, Donald Trump, bully, perpetuating*).

Due to the sparse nature of the data, the collocational profile at ± 1 in Table 2 presents weak evidence of increased negative evaluation for Sense 2 in 2016–2017. Some relevant collocates even feature only once, but they are shown because they contribute to the picture of a sense shift.⁹

9. Responding to a reviewer query, the increase of collocational span width to ± 4 barely strengthens the picture, a few of the least frequent relevant collocates only increasing by one instance or none, while the relevant collocates are inflated with sundry irrelevant ones.

Table 2. Sense 2 collocational profile for [*birther**_2016–2017]

Word	L2	L1	Total	R1	R2
<i>movement</i>			20	19	1
<i>conspiracy</i>	1		12	11	
<i>so-called</i>		6	6		
<i>theory</i>			5	2	3
<i>myth</i>			3	3	
<i>Trump's</i>	1	1	3		1
<i>attack</i>			3	3	
<i>Obama</i>		3	3		
<i>started</i>	2	1	3		
<i>pushing</i>	1		3		2
<i>claims</i>			3	3	
<i>lie</i>			2	2	
<i>lies</i>			2	2	
<i>family</i>	2		2		
<i>claiming</i>			2	1	1
<i>campaign</i>			2	2	
<i>accusations</i>			2	2	
<i>Donald</i>			2		2
<i>theorist</i>		1	2		1
<i>controversy</i>			2	2	
<i>questioning</i>			2	1	1
<i>Hillary</i>	1		2		1
<i>led</i>	2		2		
<i>rights</i>			2	2	
<i>bully</i>			1		1
<i>controversies</i>			1	1	
<i>used</i>		1	1		
<i>Barack</i>	1		1		
<i>controversial</i>		1	1		
<i>theories</i>			1		1
<i>perpetuating</i>	1		1		

Our neoseme *birtherism* reveals a reasonable pattern of sense change, given its relatively less sparse data overall, as well as its temporal and collocational separateness of senses, which facilitate the representation of social change in language by automated collocational means. Its continuing afterlife beyond the resolution of the political issue in 2011 exemplifies a characteristic of social change and its linguistic representation.

3.2 Case study 2: *Normalisation*

3.2.1 *Provenance*

The lemma *normalisation* has been an active concept and word in British English since at least 1940. It occurs throughout the corpus, evolving significantly in its meaning and reference. We condense this evolution down to four senses for ease of presentation.

The meaning ‘the acceptance as normal of politically abnormal behaviour in the Trump era’ is the target of our study. It is the lemmatized neoseme *normalisation*, also combining UK and US spelling variants *N/normalisation* and *N/normalization*, of which the former are 10 times as frequent as the latter in our corpus. A time graph in Figure 2 shows the path of all forms, cases and senses of *normalisation* across the corpus. Uses are shown to have decreased during the 1990s, and have maintained a low frequency of about one per million words until rising to four per million words and above at the end of the corpus from 2016.

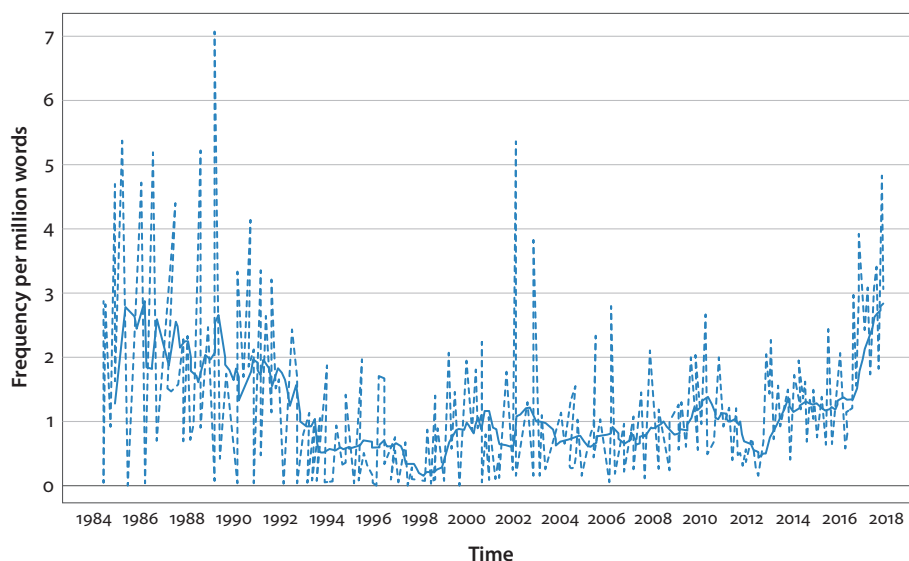


Figure 2. Time graph for 1,908 occurrences of [*normalisation|normalization*], case-insensitive

The first two definitions of the lemma *normalis/zation* are adapted from the OED entry:

- Sense 1: 1847. To make normal; to bring or return to a normal or standard condition or state.
- Sense 2: 1938. The policy of achieving and maintaining stable political and other relationships between two parties, typically countries.

Sense 2 remains throughout the corpus. It operates from a British cultural standpoint, conveying an optimistic emphasis on establishing international relations through such collocates as *diplomatic, trade, policy, relations, ties, links, talks, dialogue, countries, states, parties* (and relevant toponymy). In time, Sense 2 encompasses more ambiguous ideals, such as co-operating with regimes which are patently corrupt. In this, it presages Sense 3:

Sense 3: The assimilation of an evolving concept of normalization, and the gradual acceptance of previously controversial or taboo social issues.

In the changing world, Sense 3 shifts or extends its semantic remit, from the restoration of norms to the acceptance of controversial issues, such as abortion, freemasonry; prostitution, gay rights, etc. This extends to socio-political and psychological processes of quiet absorption of disasters as well as undemocratic, unconstitutional and even criminal behaviour, as they settle over time and become familiar. It brings Sense 3 conceptually and collocationally closer to our neoseme, Sense 4.

This latter accommodation has long existed in the US public psyche and elsewhere (viz. the 9/11 disaster), but has not been much talked about until now, according to US media sources. In 2016, Journalist Hua Hsu¹⁰ suggests that it could have begun at the Twin Tower Attack: “After 9/11, I remember talking to friends about ‘the new normal’, the way that feelings of shock just got absorbed into the inconveniences of ordinary life.” Hua Hsu continues: “Americanness is a sponge, not an ethnicity; *normalization* is a key part of how it works” (Hsu 2016). In the absence of OED definitional information for the target neoseme, Sense 4, we edit from Wiktionary:¹¹

Sense 4: A process whereby abnormal norms of behaviour are not just naturally absorbed in time, but are actively made to seem natural and wanted; this regarded as having been done through propaganda, influence, imitation and conformity.

With reference to events before and during the 2016 US presidential election campaign, we use this definition to refer to US media claims of irregular dealings by Donald Trump and his Republican supporters. Since the presidential inauguration, we expand it to encompass the US media-reported intensification of such behaviours.

This neoseme is actually recorded in the OED as early as 1864 in the form of a supposedly exemplifying citation of the following definition: “1864. *transitive*. To make normal; to bring or return to a normal or standard condition or state,” the citation being:

10. “What Normalization Means”. Contributor to *The New Yorker* from 2016.

11. <<https://en.wiktionary.org/wiki/normalization>>

- (4) 1864 *N.Y. Times* 24 Jan. These attempts to *normalize* despotism display the impotency as well as the malignity of the Executive.

The 1864 citation clearly does not fit the OED definition, since despotism cannot be “brought or returned to a normal state”, there being no such state in a civilized, democratic society, here understood as British. Instead, it captures the essence of our Sense 4 target neoseme, which is that once unthinkable behaviours are being insinuated into society as though they were normal.

It is impossible to attach an exact date to the Sense 4 neoseme with its modern reference. The tipping point may have been, as Merriam Webster.com opines, “In the wake of the 2016 presidential election, two [...] words are currently in the ether: the verb *normalize* and its related noun, *normalization*”. Journalist Gary Younge (2018) cites several cases of *normalisation* in the *Guardian* on Fri 8 Jun 2018, adding:

During the presidential election [...] mainstream conservatives, on both sides of the Atlantic, said they wanted nothing to do with him. Then he won. Those on the right, trading principle for pragmatism [...] accommodated, adapted and soon fell in line as though the line were their idea.

In our corpus, the first instance did not take long to arrive in the UK corpus: indeed, it arrived pre-election.

- (5) When media fails to call Trump rants what they truly are – hate speech – we move toward *normalizing* his racism and xenophobia. (09/2015)

3.2.2 *Assessment of challenge*

Senses 1, 2 and 3 persist throughout the corpus, while the Sense 4 neoseme only begins in 2015, thus having the task of distinguishing itself collocationally from those established senses. In addition, Sense 3 undergoes shifts of meaning, from ‘establishing stable relations’ to ‘the acceptance of a hitherto undesirable set of social circumstances’, which entail a corresponding shift of collocates, some overlapping with those of Sense 4. Furthermore, Sense 4 itself denotes an ongoing semantic path, that of accommodating the reported pre-election and post-election behaviour within the Trump regime. All these factors presage a lack of collocational discreteness which could impede the success of an automated collocational analysis for Sense 4.

The concept and word *normalisation* was originally British, but in Sense 4, it switches its reference to a primarily US concern, simply reported on in our UK news text as a foreign event. This probably limits its frequency of occurrence, a disadvantage for collocational recall. To compensate, its related forms are lemmatized, since they exceptionally all carry the same sense. Lemmatized frequencies can lead to a fuller collocational profile, as said.

3.2.3 System performance

Sense 2, concerning the establishment and maintenance of international relations, continues across the corpus, and its corresponding frequency can readily be identified collocationally. This can be assisted further by focussing on its peak use in specified corpus time chunks. The peak use is captured in the time chunks covering 1985 to 2011, and the collocational change is already apparent with a span setting of ± 2 to the node word, as seen in Table 3.

Table 3. Sense 2 collocational profile for [*normalis*_1984–2010*], case-insensitive

Word	L2	L1	Total	R1	R2
<i>relations</i>	7	1	770	370	392
<i>process</i>	51	1	81	24	5
<i>ties</i>			65	29	36
<i>diplomatic</i>		4	65	33	28
<i>situation</i>	7	2	57	1	47
<i>full</i>	1	40	42		1
<i>become</i>	11	23	36	1	1
<i>life</i>	3		35	12	20
<i>Israel</i>	5	3	34	1	25
<i>political</i>	1	16	34	7	10
<i>trade</i>	2	5	30	17	6
<i>step</i>	27		27		
<i>security</i>	4	9	25	1	11
<i>said</i>	6	5	23	2	10
<i>steps</i>	20		22	2	
<i>attempt</i>	20		21		1
<i>attempts</i>	20		20		
<i>efforts</i>	18		20	2	
<i>policy</i>	11		19	2	6
<i>way</i>	15	2	19	1	1
<i>profits</i>	1		18	7	10
<i>further</i>	5	10	16		1
<i>wants</i>	13	2	15		
<i>bilateral</i>			15	2	13
<i>things</i>	2	1	15	11	1
<i>aimed</i>	15		15		
<i>relationship</i>			14	3	11
<i>talks</i>	9		14	5	

Sense 3, concerning the evolving accommodation of both controversial and illegal behaviour, overlaps collocationally with Sense 2, though its collocates are distinguishable by virtue of the escalating seriousness of social concern they imply, including *sexual*, *bigotry*, *harmful*, *hatred*, *cruel*, *hate*, and *attacks*, as seen in Table 4. The collocational profile is confined to right-hand span +1, since this is sufficient to show relevant collocates for Sense 3, and for brevity.

Table 4. Sense 3 collocational profile for [*normalise/normalize_2017*]

Word	R1
<i>relations</i>	7
<i>monetary</i>	3
<i>defeat</i>	3
<i>gambling</i>	2
<i>conversations</i>	2
<i>sexual</i>	2
<i>Othering; US-Cuban; menstruation; deportations; internships; risk-taking; unpalatable; Spicer; breastfeeding; isolate; behaviour; bigotry; collusion; industrialised; condoms; Hanson; harmful; Nazis; surreal; hatred; clinic; classroom; cruel; inadequate; Trump's; symptoms; babies; smoking; hate; radical; Muslim; industrial; alternative; blood; sport; putting; attacks; policy; women; people</i>	1 each

Sense 4 is sufficiently frequent to be represented collocationally. Its automated profile is enriched by focussing collocationally (i) on the peak use corpus time chunks around 2016–2017 (i.e. immediately pre- and post-election), (ii) on the application of focussed syntax, *normalis/zation + of*; and (iii) on a suitable collocational span – here right-hand span –1 and 2 to the node, as seen in Table 5.

The Sense 4 collocational profile in Table 5 reflects a graver, denser set of social concerns than that for Sense 3, including *violence*, *rhetoric*, *gambling*, *Islamophobia*, *far-right*, *rape*, *abuse*, *harassment*, *sexist*, *cruelty*, *porn*, and *racism*.

In this case study, as stated at the outset, we have undermined our own, static, definition of word sense as being the sum of the collocates typically associated with a node word, by condensing several senses into Sense 3 and arguing that the same collocational profile simply increases across time. This is a pragmatic attempt to keep categories simple, while still conveying the sweep of collocational and thus semantic change.

Table 5. Sense 4 collocational profile for [*normalis/zation* + *of*_2016–2017], case-insensitive

Word	R1	R2	RTotal
<i>relations</i>	20	8	28
<i>violence</i>	4	7	11
<i>rates</i>	2	6	8
<i>interest</i>	7		7
<i>ideas</i>	2	4	6
<i>sexual</i>	5		5
<i>rhetoric</i>		4	4
<i>gambling</i>	4		4
<i>Islamophobia</i>	3		3
<i>hate</i>	3		3
<i>culture</i>		3	3
<i>far-right</i>	3		3
<i>rape</i>	3		3
<i>trade</i>	3		3
<i>abuse</i>	1	2	3
<i>harassment</i>	1	2	3
<i>sexist</i>	1	1	2
<i>politics</i>		2	2
<i>political</i>		2	2
<i>Donald</i>	2		2
<i>bad</i>	2		2
<i>Trump</i>		2	2
<i>cruelty</i>	2		2
<i>life</i>		2	2
<i>policy</i>		2	2
<i>porn</i>	1	1	2
<i>diplomatic</i>	2		2
<i>Cuban-American</i>	2		2
<i>racism</i>	2		2

3.3 Case study 3: *Cougar*

3.3.1 *Provenance*

The new sense of the noun *cougar* exemplifies a recent cultural, social-stereotyping and gender issue. *Cougar* (including *C/cougar/s*), meaning ‘an older woman who dates younger men’ denotes a sociolinguistic change in its derogatory attribution to women. Historically, it originates in the US, with a transatlantic social and semantic barrier to cross.

It is a low-frequency word, with 349 occurrences only, made up of *cougar* (213) and *cougars* (136), of which the new sense accounts for just circa 140 occurrences (manual search estimate). The similar semantics and function of *cougar* and *cougars* justify combining their profile. The OED provides the following (edited) definition of *cougar*:

Sense 1: 1774. from French *couguar*. A large feline quadruped [...] found wild in most parts of America.

The continuing corpus presence of the referent is partly due to reports on nature programmes and journalistic articles, often concerning big-cat and large animal species in general (see (6) and (7)), and hunting legislation.

(6) some 250 *cougars*, 300 grizzlies, 800 wolves and 1,000 plus black bears are shot
(02/2011)

(7) big cat hunts for mountain lion, *cougar*, puma or panther in the western US
(10/2010)

For a definition of Sense 2, we turn to Oxford Dictionaries:¹²

Sense 2: (informal) an older woman seeking a sexual relationship with a younger man.

The origin of this sense seems to be a Canadian dating website called *Cougardate.com*, started in 1999 by two women, who took the name from the Canadian NHL Hockey team, who labelled older female fans “Cougars”,¹³ implying some uncharacteristically masculine traits in the psyche and behaviour of women. The tipping point for the popularity of the concept and the neoseme seems to be a 2002 book *Cougar: A Guide for Older Women Dating Younger Men*, by Valerie Gibson (Gibson 2002).

However, *cougar* did not reach our UK corpus until 2007. In explaining the process of diffusion which drove the transatlantic shift of the neoseme, one looks to word of mouth (albeit incalculable), social media, dating websites, news articles, books, films and TV series, notably *Sex and the City* (1998–2004), and *Cougar Town* (2009–2015), which both came to the UK.

The first occurrence of *cougar/s* as a neoseme in our news corpus appears in June 2007, shown in (8):

(8) she is what Maxim [UKUS] magazine calls a “*cougar*”: an older woman with a weakness for attractive younger men like him.
(06/2007)

12. <<https://en.oxforddictionaries.com/definition/cougar>>

13. <<https://www.quora.com/How-did-women-who-seek-younger-men-get-to-be-called-cougars>>

The emergence of Sense 2 of *cougar/s* in our UK news text is illustrated by a time graph of instances specified as collocating with *woman/women* within span ± 30 , amounting to 75 occurrences, as shown in Figure 3.

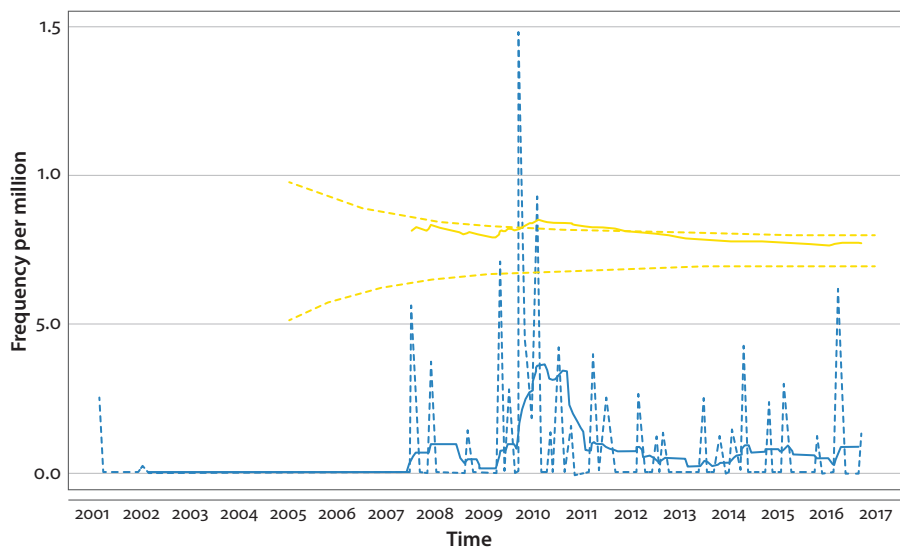


Figure 3. Time graph of frequency for [*cougar/s*] in Sense 2, with growth from mid-2007

The time graph in Figure 3 shows the new sense of *cougar/s* emerging in 2007–2009, rising in 2009–2012, reaching peak use in 2009–2011, falling to smaller peaks in 2014 and 2016, but remaining stably rare thereafter, at around 0.1 per million words.

References are from early on made to UK contexts. As with *birtherism* and *normalisation*, its presence endures in the corpus beyond its peak use, this time due to newspapers trading in social stereotyping and Hollywood-style celebrity behaviour and scandal.

3.3.2 Assessment of challenge

Since *cougar/s* used of women is a metaphorical extension of *cougar* used of the wild animal, some collocational duplication can be expected between the two senses which undermines the distinctiveness of the collocational profiles, for instance, *hunt*, *prey*, *predator(y)*, *prowl*. This, together with the sparseness of the corpus data for *cougar/s* in Sense 2, is predicted to make it difficult to isolate and reveal the semantic change automatically. For Sense 2, the combining of profiles for *cougar* and *cougars* is likely to boost frequency counts and so improve the results.

3.3.3 System performance

Despite the sparseness of the collocational data for Sense 2, a collocational analysis is achieved by a series of means: firstly, by focussing on the subset of the corpus time chunks in which the peak use occurs, which is 2009–2011/12. Due to the rarity of the neoseme (circa 142 occurrences), the results are nevertheless rather sparse, and the collocates not all relevant. The relevant collocates are printed in bold type, in the collocational profile for *cougar/s*, the neoseme, in Table 6.

Table 6. Sense 2 collocational profile for [*cougar/s_2009–2012*] span ± 4

Word	L4	L3	L2	L1	Total	R1	R2	R3	R4
<i>older</i>		1		1	10	2	2	3	1
<i>term</i>				7	9			2	
<i>woman</i>			1		8		1	2	4
<i>such</i>			4		5				1
<i>women</i>	1		1		4		1	1	
<i>wolves</i>	1		1		4		1		1
<i>mountain</i>		1	1		3		1		
<i>men</i>			1	1	3				1
<i>young</i>					3				3
<i>prowl</i>					3			2	1

Secondly, a better result is achieved by extending the corpus timeline from 2009–2012 to 2009–2016, as seen in Table 7. This allows some of the later US TV and film productions based on the adventures of “cougar” women to be reflected in the corpus data for system processing. Thirdly, the search string has also been extended to include upper-case *Cougar*, in the singular, to allow in film and TV titles but exclude rare references to sports teams (e.g. *Vancouver Cougars*; *Keithley Cougars Rugby*). Fourthly, a *Z-score* (lower significance threshold around ≥ 2) has additionally been applied, but it happens to make no change to the collocational significance ordering at this level. Finally, a left-right collocation span of (3)–(4) also works better.

The neoseme *cougar/s* is a rarely occurring item, which overlaps with a previous sense which endures throughout the corpus, but which nevertheless can be induced to furnish an automated collocational profile, by means of extended timelines and judicious use of search string case variants.

Table 7. Sense 2 collocational profile of [*cougars*|*Cougars_2009–2016*] case-insensitive

Word	L4	L3	Total	R3	R4	ZScore
<i>Living</i>	1	8	11	1	1	9.94
<i>Sky1</i>			7		7	6.00
<i>older</i>	1	1	7	4	1	5.75
<i>woman</i>			6	2	4	4.48
<i>Courteney</i>	1		5	1	3	4.00
<i>Cox</i>			5	2	3	3.97
<i>John</i>			5	5		3.10
<i>prowl</i>			4	3	1	3.00
<i>Bridges</i>			4		4	2.99
<i>gags</i>		4	4			2.99
<i>Courtney</i>		2	4	1	1	2.99
<i>adolescent</i>	4		4			2.99
<i>bears</i>	1		4	3		2.97
<i>dating</i>	2	2	4			2.97
<i>Moore</i>	1		4	1	2	2.93
<i>animals</i>	2	2	4			2.90
<i>Kevin</i>			4	4		2.89
<i>term</i>			4	4		2.78
<i>pm</i>			4	2	2	2.24
<i>Philipps</i>			3	2	1	2.00
<i>Established</i>		1	3	2		2.00
<i>wolves</i>	1		3		2	1.99
<i>predatory</i>			3	2	1	1.99

3.4 Case study 4: *Snowflake*

The target neoseme *snowflake*, Sense 2 in our study (see below), denotes, like *cougar*, a recent addition to social stereotyping. It means ‘someone with an exaggerated sense of entitlement and oversensitivity to criticism’. A closely associated Sense 2a adds a political dimension to the same definition (see below). Senses 2 and 2a are both primarily conveyed by the lower-case singular form *snowflake*. Upper-case *Snowflake* risks retrieving company names; plural *snowflakes* risks collision with Sense 1, ‘small masses of snow’.

3.4.1 *Provenance*

The OED provides the following definition:

Sense 1: 1734. One of the small masses in which snow commonly falls.

The word *snowflake* is heavily polysemous, with the many senses following Sense 1 in the OED all making metaphorical allusion to the physical qualities of snow in its literal sense; in appearance, structure, fragility or colour. We shall focus only on those relevant to our study.

The OED has no entries for the more recent neosemes for *snowflake*. Wiktionary provides the following definition:

Sense 2: 1996. (US slang, derogatory) Someone who believes they are as unique and special as a snowflake; someone hypersensitive to insult or offense, especially a young person with politically correct sensibilities.

According to Wikipedia, *snowflake* is claimed by Chuck Palahniuk to have originated from his 1996 novel *Fight Club*, where a line reads “You are not special. You are not a beautiful and unique snowflake. You’re the same decaying organic matter as everything else” (Evening Standard 2017).

An article in *The Australian* in 2016 by Helen Rumbelow (Rumbelow 2016) says that our Sense 2 neoseme began in the U.S. in 2015, following a student/faculty confrontation at Yale University. Swiftly, *snowflake* widened its reach, used by tabloids to express generic disdain for young people.

The term “Generation Snowflake”, or its variant “Snowflake Generation”, probably came to the UK in 2016 following the publication of Claire Fox’s book entitled *I Find That Offensive!* (Fox 2016). Soon after, in 2016, the neoseme appeared in The UK *Daily Telegraph*, where Bryony Gordon,¹⁴ also a Briton, writes:

My favourite title for today’s young folk, though, is the *Snowflake Generation*. I heard it on the television this week, as [...] a student had been threatened with violating university ‘safe space’ rules, simply by raising her hand. (Gordon 2016)

Snowflake remains a rare neoseme in the British English corpus, although 42 out of the 45 instances of Sense 2 are found in British sources. The epithet has not caught on significantly, whether in reportage about US or even UK youth. *Snowflake generation* has aroused the UK media’s attention, though it is still rare (25 occurrences) in our 2017-ending corpus. In 2016, the *Collins English Dictionary* added *snowflake generation* to its ‘Words of the Year’. The term referred to ‘millennials’,

14. <<https://www.telegraph.co.uk/opinion/2016/04/08/bryony-gordon-why-i-feel-sorry-for-the-snowflake-generation/>>

then meaning youth born from 2000 on; by 2017/8, also to youth born from 1985 or 1990 – millennials who were adult post-2008 financial crash (see Hosie 2017).¹⁵

A recent elaboration of Sense 2 introduces a new Sense 2a, which features in no dictionary as yet. It is not an entirely separate sense, but rather a narrowing of reference which has crept into US political discourse in 2017, and even into UK ministerial rhetoric. We propose the following definition (derived from Wikipedia):

Sense 2a: A pejorative term used of liberals by Trump and right-wing US (and UK) politicians, to ridicule them for disagreeing with Republican tactics; reportedly directed at President Trump for his poor reaction to criticism.

An article by US journalist Mark Abadi¹⁶ (Abadi 2017) explains that:

2017 saw the word “snowflake” gain traction as a pejorative used by conservatives to insult liberals perceived as being overly sensitive [...] The new usage of the word became especially popular among [...] supporters of President Donald Trump.

Jessica Roy says, in the *Los Angeles Times*, that “the alt-right in the United States pejoratively describes most liberals and those protesting against Donald Trump as ‘snowflakes’, short for ‘special snowflake’” (Roy 2016). John Paul Brammer, a US journalist, broadens the definition in the UK *Guardian* on January 16, 2017 (Brammer 2017):

The word [...] has been recast as meaning almost any kind of upset, no matter how minor [...] It’s ironic, then, that the public figure who most reliably projects those qualities at the moment is Donald Trump. America is soon to inaugurate its first Snowflake-in-Chief.

During 2017, the occasional use of the Sense 2a political reference to *snowflake* creeps into the UK corpus with UK references:

- (9) *Snowflake* is a social media term, particularly favoured on the right, for people who are easily offended. (01/2017)
- (10) Gove’s ‘*snowflake*’ tweet is symptomatic of the nastiness that’s ruining politics. He tweeted: “People ‘offended’ by the foreign secretary’s comments today are humourless, deliberately obtuse, snowflakes”. People who thought the foreign secretary might consider watching his words during delicate EU negotiations are thin-skinned idiots. (01/2017)

15. See Rachel Hosie in May 2, 2017, *Independent*: “The millennial divide: How a generation can be split into two very different halves”.

16. <<https://www.businessinsider.com/words-changed-meaning-2017-12>>

- (11) The *Snowflake* Tendency has even begun to infect political discourse in Scotland.
(03/2017)
- (12) Me, a left-wing *snowflake*, fully signed up to killing comedy and suppressing
free speech.
(12/2017)

3.4.2 Assessment of challenge

Snowflake is heavily polysemous, which bodes ill for the production of clear, discrete collocational profiles for the target neoseme. Sense 2, our neoseme, occurs only 45 times in the corpus. This makes for a very sparse collocational profile, if any. Lemmatization would be a means of enriching the data available, though *snowflakes* is ambiguous between Sense 1 and 2, and the senses overlap timewise.

3.4.3 System performance

For Sense 2, denoting ‘oversensitive young people’, it is possible to produce a short collocational profile, as seen in Table 8, by applying a Z-score (significance threshold down to ≥ 2) which promotes and groups most relevant collocates in bold type; (*liberals* is marked with an asterisk, since it appears prematurely, belonging to Sense 2a; *broflake*, *coddling*, and *oversensitive* will appear later in Sense 2a).

Table 8. Sense 2 collocational profile for [*snowflake_Nov2016–7*] lowercase, low Z-score

Word	L2	L1	Total	R1	R2	ZScore
<i>generation</i>	1	6	25	18		23.69
<i>special</i>		8	8			6.81
<i>unique</i>		5	5			3.96
<i>falling</i>		1	5	3	1	3.94
<i>liberals*</i>		1	3	1	1	2.00
<i>whiny</i>	1	1	2			1.00
<i>millennials</i>			2	2		1.00
<i>tendencies</i>			2	2		1.00
<i>fielding</i>			2		2	1.00
<i>catching</i>	2		2			0.99
<i>Sensitive</i>		2	2			0.99

For Sense 2a, if the Sense 2 profile is allowed to continue downwards into hapax occurrences, it eventually reaches the collocates belonging to the politically-referenced Sense 2a. These can be seen in Table 9: that is, *cuck* (right-wing insult for liberal men), *feminazi* (right-wing insult for feminist women), *Trumpism*, and *alt-right*, (but missing *liberals*; see Table 8).

The Sense 2a profile even contains five UK-referenced political collocates (marked with an asterisk in Table 9); they overlap timewise and collocationally, and are very infrequent.

Table 9. Sense 2a collocational profile for [*snowflake_Nov2016–2017*] lowercase, 0.00 Z-score

Word	L2	L1	Total	R1	R2	ZScore
<i>broflake</i>	1		1			0.00
<i>Medium's</i>		1	1			0.00
<i>cuck</i>			1		1	0.00
<i>Nakaya</i>			1	1		0.00
<i>remoaner*</i>		1	1			0.00
<i>cry-baby*</i>			1	1		0.00
<i>feminazi</i>			1	1		0.00
<i>stylebook*</i>		1	1			0.00
<i>coddling</i>	1		1			0.00
<i>over-sensitive</i>		1	1			0.00
<i>Trumpism</i>	1		1			0.00
<i>seditious</i>		1	1			0.00
<i>bauble</i>			1	1		0.00
<i>claptrap</i>			1	1		0.00
<i>bulldozing*</i>			1		1	0.00
<i>silver-haired</i>		1	1			0.00
<i>spendthrift</i>		1	1			0.00
<i>Cheer</i>	1		1			0.00
<i>feathery</i>			1		1	0.00
<i>alt-right</i>		1	1			0.00
<i>knits</i>			1	1		0.00
<i>redtop*</i>	1		1			0.00

As with the semantic evolution of *normalisation*, a catalyst on the *snowflake* semantic path from the ‘over-sensitive youth’ of Sense 2 to the ‘over-sensitive liberal politician’ of Sense 2a is reported to be the US president himself, both as addresser and addressee, as suggested in (13) to (17) representing occurrences of [*snow#flake**], with filter +*Trump** in sentence.

- (13) Trump was repeatedly labelled a *snowflake* earlier this month (11/2016)
- (14) Trump said the theatre should [...] be a safe space; the irony not lost on commentators, who called him the most special *snowflake* of all. (11/2016)
- (15) an ideological split between Trump and his climate-change-denying [...] moronic supporters and *snowflakes* with their [...] loser liberalism (01/2017)
- (16) Trump is, of course, king of the *snowflakes* (02/2017)
- (17) With Trump, Yiannopoulos became a gleeful scourge of liberal *snowflakes* (02/2017)

This assertion appears too rarely in the UK corpus to support a collocational profile.

3.5 Case study 5: *Ghosting*

3.5.1 *Provenance*

The lemmatized neoseme *ghosting*, Sense 2 in our study (see below), denotes a technologically newly-enabled social process in online-dating culture, complementing the range of recent social changes illustrated in previous case studies. The term itself is not pejorative, but denotes a socially unacceptable if widespread practice, experienced and perpetrated by all genders, of ending a personal relationship by abruptly ceasing communication. Though it probably began in the US electronic gaming culture, it soon became globally established.

The earliest OED entry for the word *ghost*, as a verb, dates from the 19th century, edited as follows:

Sense 1: 1833. To move like a ghost; to steal away like a ghost.

The OED and other sites together provide an extensive record of its polysemy. Subsidiary senses all seem to draw on the analogical or metaphorical potential of the 17th century practice of acting like a ghost, being ghost-like, to denote activity across the social spectrum. Domain-coded in Table 10 are some senses cumulatively occurring in the corpus sufficiently to hamper the automatic identification of sense change.

Table 10. Inventory of senses recorded in major lexicons for *ghost/ghosting*

Year	Domain	Sense
1920	Identity theft	A person takes on the identity of a deceased person
1922	Literature	Writing something as a ghost-writer
1990	Sport	Strategic moves on pitches and courts during football and tennis matches
1991	Psych/mental	Thoughts staying or swirling around in the mind
1991	Prison fraud	Moving disruptive prisoners to other prisons to avoid inspection
1992	Music	Playing with differing hand pressure on the piano
1994	Visual technology	A replica of the transmitted image superimposed over the main image
1996	Dance choreography	Gay dance technique whereby two dancers echo each other's moves
2006	US military fraud	Secret movement of detainees abroad; detention in foreign prisons
2012	UK municipal fraud	Unrecorded school exclusions of disruptive pupils
2017	Film technology	Use of a double in filming, or to insert unavailable or deceased actor
2018	Employment	An interviewee is given to think s/he has a chance, but doesn't hear back

Of these interim senses, the two most apparent throughout the corpus are the 1990 'strategic sporting moves', followed by the 1922 'ghost writing'.

For Sense 2, our neoseme, a definition is edited from Wiktionary and Wikipedia, as follows:

Sense 2: 2011. A method of ending a personal relationship by stopping any contact with the other party and not providing an explanation, increasingly via social media and online dating apps.

The practice of ending a relationship by silence is time-honoured (cf. e.g. *she dropped him*), but the online dating context is new. Wikipedia¹⁷ reports on probable tipping points in its dissemination:

The term is attested since at least 2006, in the context of online exchanges, and became popular by 2015 through numerous articles on high-profile celebrity relationship dissolutions [...] It has been the subject of numerous articles and discussions [...] It was included in the *Collins English Dictionary* in 2015.

In the case of our Sense 2 neoseme, the transition from real-world change to language change is fairly seamless, with the first occurrence appearing promptly in the UK corpus (see (18)), albeit in mention rather than use:

- (18) ghosting: to mean ending a relationship by ignoring communications
(05/11/2015)

3.5.2 *Assessment of challenge*

Our lemmatized neoseme *ghosting* comprises circa 439 occurrences. The issue is whether this low count is sufficient to withstand the barrage of irrelevant senses which precede it and produce an automated collocational analysis. An unbroken stretch of output from our corpus for 2016–2017, a period which might be expected to focus in on the ‘dating’ neoseme, is coded in (19)–(29) to show just how interspersed the various senses are (cf. Literature, Relationships, Sport, Prison, Social Behaviour, Cinema, Music).

- (19) Writers who have spilled the beans about *ghosting* include Jennie Erdal
(07/2016) Literature
- (20) Swiping left, *ghosting*, [...] dating has never been so brutal.
(07/2016) Relationships
- (21) I was having *ghosting* images, halos, glare. (08/2016) Visual
- (22) he was *ghosting* into Southampton’s area to ping a shot at Fraser Forster
(08/2016) Sport
- (23) The pulsating Swayze is particularly poignant: it’s about *ghosting*.
(08/2016) Relationships

17. <[https://en.wikipedia.org/wiki/Ghosting_\(relationships\)>](https://en.wikipedia.org/wiki/Ghosting_(relationships)>)

- (24) the government intends to [...] target a group of prisoners with *ghosting* as a strategy (08/2016) Prison
- (25) the creator [...] *ghosting* away from Ryan Mason before feeding Lallana's run (09/2016) Sport
- (26) love have been replaced with modern iterations (saucy pictures and *ghosting*) (09/2016) Relationships
- (27) instead of *ghosting* out of the room with you they stop to farewell everyone (06/2017) Social behaviour
- (28) Peter Cushing via *ghosting* in Rogue One: A Star Wars Story If ever there (06/2017) Cinema
- (29) Dreamy Caramello boasts breathy, sweet vocals *ghosting* over rich [...] electronic textures (11/2017) Music

Attempts at automatic search reveal that *ghosting* in 2015–2017 is, as anticipated, primarily swamped by instances of 'sporting movement', followed by those of 'ghostwriting' (246 occurrences just of *ghost** + *by*). The sporting sense could be manually down-weighted by eliminating its largely prepositional right-hand collocates (*in*, *along*, *behind*, *between*, *past*, etc), and the 'ghostwriting' sense by elimination of an inventory of writing types (*book*, *memoirs*, *autobiography*, *diary*, *books*, *column*, *soaps*, etc). Manual manipulation is useful for confirming suspicions (and sometimes incidentally revelatory), as well as being necessary given the user-defined design of our system.

Thus, the neoseme *ghosting* (dating) threatens to be a case where our supposition was right, in that the automated system based on collocational frequency alone cannot distinguish the relevant contexts from the myriad ambiguous or irrelevant ones.

3.5.3 System performance

With insufficient collocates to work with, and a forest of interwoven irrelevant senses, we make a final attempt to design a search string which will optimize the chances of capturing the Sense 2 neoseme in an automated collocational change profile. A number of measures are applied:

- i. Focussing on the relevant time chunks of corpus data, namely 2016 to 2017;
- ii. Expanding the span to the maximum of ± 4 ; and
- iii. Tailoring the search string to require *ghosting* to be a verb, immediately followed by an object noun or pronoun, or by the collocate *by*, to catch passive constructions. This will not ensure the retrieval of all relevant collocates, but perhaps retrieve a viable threshold.

Surprisingly in the circumstances, as many as 14 out of 31 of the retrieved collocates prove, when manually checked, to relate to *ghosting* in Sense 2 (see the words emboldened in Table 11).

Table 11. Sense 2 collocational profile for [*ghosting/ghosted*(V)+N/P/ by_2016–7]

Word	L2	L1	Total	R1	R2
<i>way</i>			2	1	1
<i>writer</i>	1		2		1
<i>email</i>		1	1		
<i>Guilty</i>	1		1		
<i>crews</i>			1		1
<i>man</i>		1	1		
<i>provide</i>			1		1
<i>left</i>		1	1		
<i>Reagan's</i>			1	1	
<i>halos</i>			1	1	
<i>masters</i>	1		1		
<i>Personally</i>			1		1
<i>disappearance</i>	1		1		
<i>messages</i>			1		1
<i>glare</i>			1		1
<i>effectively</i>		1	1		
<i>produce</i>	1		1		
<i>Cushingstyle</i>		1	1		
<i>non-relationship-based</i>	1		1		
<i>hacks</i>			1		1
<i>seldom</i>			1	1	
<i>ex</i>			1		1
<i>20</i>	1		1		
<i>literary</i>	1		1		
<i>Daily</i>			1	1	
<i>X</i>	1		1		
<i>almost</i>			1	1	
<i>Mirror</i>			1		1
<i>gave</i>			1		1
<i>excites</i>			1		1
<i>Apple</i>			1	1	
<i>text</i>			1	1	

The purpose in selecting the *ghosting* neoseme was to show how a low-frequency neoseme fares in a vast array of polysemic competitors. We show that it can be induced to differentiate itself weakly. In (30) to (33), it conversely serves to illustrate

how this inducement can permit the inclusion of collocates which are in fact semantically irrelevant.

The criterion of collocational significance applied in our automated analysis allows for cross-sentence-boundary collocation within a user-specified span, normally up to ± 4 to the node.¹⁸ While this may be significant, as Greenbaum (1970: 11) showed with lemmas *stamp* and *collecting*, it does not hold that all words occurring within the span are actually involved in a significant semantic relationship with the node. This is particularly likely when the data are sparse, and singly-occurring words are included. In (30)–(33), we see that some collocates are not intrinsic to the neoseme *ghosting*.

- (30) no one referred to his disappearance as *ghosting*. It's *almost* bizarre (02/2016)
- (31) Letters and unrequited love have been replaced with modern iterations (*saucy* pictures and *ghosting*). (09/2016)
- (32) *Oh* so Gary *ghosted* you? No way (05/2017)
- (33) I had no idea that non-relationship-based email *ghosting* was a thing. *Personally*, I'm not up to it. (12/2017)

This case study of *ghosting* further demonstrates that Z-score significance statistics are not reliable for use at the hapax level, something self-evident. While promoting some genuine collocates such as *non-relationship-based*, *swiping*, *cushioning*, *reciprocate*, *tearful*, *ex*, and *disappearance*, Z-score measures also offer span ± 4 words unrelated to the node *ghosting*, as in (34).

- (34) Guilty of *ghosting*, breadcrumbing or cushioning? *Looks* like your secrets are out. (05/2017)

4. Discussion

The earliest sense of all the base forms is recorded in the OED or other lexicographic sources as being of British English origin, while their neosemes originate abroad, primarily North America, before being quoted and/or assimilated into our British English news corpus. Consonant with the unsettled state of the transatlantic nations, it transpires that all neosemes denote states, behaviours or processes which are controversial, with negative connotations. With the exception of *cougar* and *ghosting*, which can be used confessionally, the neosemes are not used self-referentially, but

18. During the AVIATOR project, the team compared the results of different collocational spans, and established that the default for English was 4 words to the left and right of the node, with larger spans introducing noise; French requires a span of 6.

negatively of others, often political adversaries. These neosemes appear and advance from the 1990s into the 2000s and on until the transatlantic political upheaval and the era of Trump.

4.1 Challenges

The first challenge is that our neosemes are all of low frequency. For the most recent in our corpus, *ghosting* (2015), *normalisation* (2016) and *snowflake* 2a (2017), this may be because they are all still in use at close of corpus in December 2017. For earlier neosemes, *cougar* (2007) peaked in 2009–2012, became a cliché, and continues at a low frequency level of 0.25 occurrences per million words. Similarly, *birtherism* (2009), which peaked in 2009–2011, continues at around 0.1 and 0.4 words per million, but peaked again in 2016.

Another challenge is that our neosemes, whilst etymologically rooted in British English, all refer for the most part to issues originating or popularized in US news media, and are only marginally employed in UK cultural reference. This in turn leads to a transatlantic delay in their UK appearance, and thus to a low frequency of occurrence.

A further challenge is that neosemes are, by definition, polysemous items, usually metaphorical extensions of literal roots. Metaphor depends for comprehensibility on allusion to the literal root sense by partial comparison, which is achieved by the selective retention of strategic collocates from that root word. This leads to collocational overlap. Heavily polysemous words entail a series of neosemes emerging over time, each partially alluding to the root sense, as with *normalisation*, *snowflake*, and *ghosting*. This potentially exacerbates the problem of collocational overlap.

4.2 Measures shown to allow or enhance system performance

In dealing with our low-frequency neosemes, where none has crossed the 1,000 occurrence frequency threshold (i.e. an overall frequency of 1 per million words or above, our observed threshold for statistical viability), it has generally been necessary to establish measures to improve both recall and precision for each analysis.

The first measure is recall: to boost the volume of available collocation data, we can

- i. extend the collocational span to a maximum of ± 4 (longer does not generally achieve anything); however, while real collocational relations can occur at this span, so can semantically unrelated ones (Greenbaum 1970: 11), particularly towards the hapax level of word occurrence, or when the collocational span is allowed to cross sentence boundaries (see *ghosting*);

- ii. combine, or if linguistically appropriate, lemmatize the derived and inflected forms of the neoseme; and
- iii. allow upper- and lower-case typographical variants.

The second measure is precision: to increase the likelihood of pinpointing relevant collocational data, we can

- i. restrict the collocational span to the most relevant positioning within ± 4 (e.g. *cougar* analysis optimal at span ± 3 and 4);
- ii. restrict search to corpus time chunks recording peak use to find more relevant collocates for the neoseme;
- iii. restrict the selection of derived and inflected forms of a neoseme to the most relevant;
- iv. use Z-score statistics to promote or demote collocational profiles (less reliable if neoseme is so rare as to attract a Z-score approaching zero);
- v. focus the search string with lexical or syntactic pattern specification; and
- vi. tailor the search by manually weighting the search string with required collocates (a method neither automatic nor empirical; but it may confirm what is known or predicted).

4.3 Sociolinguistic insights gained in the study

There is an unpredictable pattern of diffusion which takes place between a social change and its emergence in the language as a neoseme. There is also a delay involved before that social change and/or neoseme reaches society, particularly the major language disseminators (journalists, politicians); and a further delay involved in its reaching the UK. The emergence of language change to accompany social change involves complex factors, and typically a tipping point.

Neosemes in this study are still active in the corpus, so it remains to be seen how they develop. Those specific to US language change, particularly fleeting change, may partially but not totally assimilate into UK language or culture. For instance, in 2019, *snowflake* begins to appear with distinct UK social and political reference; *cougar* in (particularly tabloid) reference to UK celebrity relationships; *ghosting* still denotes cutting romantic contact, but has gained a wider remit, to include simply 'withdrawing socially', or 'staying at home', in UK youth vernacular.

Once past their peak use, reference or popularity, many neosemes manifest a protracted afterlife (see *birtherism* and *cougar*), due to a journalistic or political desire to keep a topic alive, to retrospectively analyze and take stock of past events, to create a 'retro' stylistic effect, and so on.

5. Conclusion

In this study, our starting point has been contemporary real-world social change, and its linguistic expression by means of neosemes. The study has briefly outlined the history and pattern of diffusion of each neoseme, noting where possible the tipping point of popularization.

The study has analyzed five recent abstract neosemes, denoting social processes (e.g. *birtherism*, *normalis(z)ation*, *ghosting*), and social stereotypes (e.g. *cougar*, *snowflake*), as they join the changing discourse of American and UK society. The task was to test whether the automated system detects the sense changes through collocational analysis. The chapter has examined each neoseme in a case study, focussing on the challenges facing them, due primarily to their sparseness of occurrence, as new formations, or to the degree of overlap in collocation which they share with related words, typically previous senses.

Regarding the performance of the WebCorpLSE automated collocation analysis software, what has been revealed is that it achieves considerable success given its conceptual simplicity. Its ability to reflect language change depends on the presence in the corpus of sufficient collocational data, however. To reflect social change, it depends on a critical threshold of frequency and promptness with which social change is talked about and recorded in articles and quotations which make their way into the news corpus. Its default functionality has been shown, as well as various measures which can be undertaken to allow or enhance its performance in teasing out collocational profiles from sparse data.

As said previously, the WebCorpLSE software is currently being made pro-active in the area of word frequency change identification, and similar enhancement is in prospect for semantic change.

References

Corpora and tools

- ACRONYM = Automatic Collocational Retrieval of NYMS project, 1994–1997, Liverpool University. <<http://rdues.bcu.ac.uk/acronym.shtml>>
- APRIL = Automatic Prediction of Innovation in Language project, 1997–2000, Liverpool University. <<http://rdues.bcu.ac.uk/april.shtml>>
- AVIATOR = Analysis of Verbal Interaction & Text Retrieval project, 1990–1993, Birmingham University. <<http://rdues.bcu.ac.uk/aviator.shtml>>
- WebCorpLSE = WebCorp Linguist's Search Engine tools. <<http://wse1.webcorp.org.uk/>>

Other references

- Abadi, Mark. 2017. "'Snowflake,' 'unicorn,' and all the other words that changed meaning in 2017". *Business Insider*. <<https://www.businessinsider.com/words-changed-meaning-2017-12?international=true&r=US>> (12 November 2019).
- Bauer, Laurie. 1983. *English Word-Formation*. Cambridge: CUP.
- Barbaro, Michael. 2016. "Donald Trump Clung to 'Birther' Lie for Years, and Still Isn't Apologetic". *Nytimes.com*. <<https://www.nytimes.com/2016/09/17/us/politics/donald-trump-obama-birther.html>> (12 November 2019).
- Brammer, John P. 2017. "America: behold, your Snowflake-in-Chief". *The Guardian*. <<https://www.theguardian.com/commentisfree/2017/jan/16/snowflake-in-chief-donald-trump>> (12 November 2019).
- Brezina, Vaclav, McEnery, Tony & Wattam, Stephen. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2): 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>
- Boussidan, Armelle & Ploux, Sabine. 2011. Using topic salience and connotational drifts to detect candidates to semantic change. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*, Johan Bos & Stephen Pulman (eds), 315–319. <<http://www.let.rug.nl/bos/pubs/BosPulman2011.pdf>> (3 February 2020).
- Cabré, Teresa & Nazar, Rogelio. 2012. Towards a new approach to the study of neology. In *Neologica 6*, John Humbley & Jean-François Sablayrolles (eds), 63–80. Paris: Editions Classiques, Garnier.
- Cartier, Emmanuel. 2016. Neovéille, système de repérage et de suivi des néologismes en sept langues. In *Neologica 10*, John Humbley & Jean-François Sablayrolles (eds), 101–131. Paris: Editions Classiques, Garnier.
- Evening Standard. 2017. *Fight Club author Chuck Palahniuk takes credit for 'snowflakes'*. <<https://www.standard.co.uk/news/londoners-diary/londoners-diary-chuck-palahniuk-i-coined-snowflake-and-i-stand-by-it-a3448226.html>> (12 November 2019).
- Fox, Claire. 2016. *I Find That Offensive!* London: Biteback Publishing.
- Gibson, Valerie. 2002. *Cougar: A Guide for Older Women Dating Younger Men*. Toronto: Key Porter Books.
- Gordon, Bryony. 2016. "I feel sorry for the poor ickle millennials". *The Telegraph*. <<https://www.telegraph.co.uk/opinion/2016/04/08/bryony-gordon-why-i-feel-sorry-for-the-snowflake-generation/>> (12 November 2019).
- Greenbaum, Sidney. 1970. *Verb-Intensifier Collocations in English: An Experimental Approach*. The Hague: Mouton.
- Hosie, Rachel. 2017. "Millennials can be split into two distinct groups and it all comes down to one date". *The Independent*. <<https://www.independent.co.uk/life-style/millennial-divide-generation-split-old-young-financial-crash-internet-smartphones-a7704021.html>> (12 November 2019).
- Hsu, Hua. 2016. "What Normalization Means". *The New Yorker*. <<https://www.newyorker.com/culture/cultural-comment/what-normalization-means>> (12 November 2019).
- Kehoe, Andrew & Gee, Matt. 2009. Weaving Web data into a diachronic corpus patchwork. In *Corpus Linguistics: Refinements and Reassessments* [Language and Computers 69], Antoinette Renouf & Andrew Kehoe (eds), 255–279. Amsterdam: Rodopi. https://doi.org/10.1163/9789042025981_015

- Merriam-webster.com. (n.d.). "The New 'Normalize'". <<https://www.merriam-webster.com/words-at-play/normalize-normalization-meaning>> (12 November 2019).
- Mollet, Eugene, Wray, Alison & Fitzpatrick, Tess. 2011. Accessing second-order collocation through lexical co-occurrence networks. In *The Phraseological View of Language: A Tribute to John Sinclair*, Thomas Herbst, Susan Faulhaber & Peter Uhrig (eds), 87–122. Berlin: Mouton de Gruyter.
- Oxford English Dictionary Online*. 2000. Michael Proffitt (ed.), 3rd edn. Oxford: OUP. <<http://www.oed.com/>> (7 November 2019).
- Renouf, Antoinette. 1993. A word in time: First findings from dynamic corpus investigation. In *English Language Corpora: Design, Analysis and Exploitation*, Jan Aarts, Pieter de Haan & Nelleke Oostdijk (eds), 279–288. Amsterdam: Rodopi.
- Renouf, Antoinette. 1996. The ACRONYM Project: Discovering the Textual Thesaurus. In *Synchronic Corpus Linguistics: Papers from the 16th International Conference on English Language Research on Computerized Corpora (ICAME 16)*, Carol Percy, Charles Meyer & Ian Lancashire (eds), 171–187. Amsterdam: Rodopi.
- Renouf, Antoinette. 2007. Tracing lexical productivity and creativity in the British media: The Chavs and the Chav-nots? In *Lexical Creativity, Texts and Contexts* [Studies in Functional and Structural Linguistics 58], Judith Munat (ed.), 61–89. Amsterdam: John Benjamins.
- Renouf, Antoinette. 2010. Identification automatique de la néologie lexicologique et sémantique: Questions soulevées par notre méthode. In *Actes del I Congrés Internacional de Neologia de les Llengües Romàniques*, Maria Teresa Cabré, Ona Domènech, Rosa Estopà, Judit Freixa & Mercé Lorente (eds), 129–141. Barcelona: Institut Universitari de Lingüística Aplicada.
- Renouf, Antoinette. 2013. A finer definition of neology in English: The lifecycle of a word. In *Corpus Perspectives on Patterns of Lexis* [Studies in Corpus Linguistics 58], Hilde Hasselgård, Jarle Ebeling & Signe Oksefjell Ebeling (eds), 177–207. Amsterdam: John Benjamins.
- Renouf, Antoinette. 2014. Semantic neology: The challenges for automatic identification. In *Neologica 8*, John Humbley & Jean-François Sablayrolles (eds), 185–200. Paris: Editions Classiques, Garnier.
- Reutenauer, Coralie. 2012. Propositions pour la détection automatique de la néologie sémantique. *Cahiers de Lexicologie* 100: 137–160. Paris: Editions Classiques, Garnier.
- Roy, Jessica. 2016. "Analysis: 'Cuck', 'snowflake', 'masculinist': A guide to the language of the 'alt-right'". *Los Angeles Times*. <<https://www.latimes.com/nation/la-na-pol-alt-right-terminology-20161115-story.html>> (12 November 2019).
- Rumbelow, Helen. 2016. "Generation snowflake: Why millennials are mocked for being too delicate". *The Australian*. <<https://www.theaustralian.com.au/world/the-times/generation-snowflake-why-millennials-are-mocked-for-being-too-delicate/news-story/2f885d016af1a881bff92b-69282fd88e>> (12 November 2019).
- Rutenberg, Jim. 2008. "The Man Behind the Whispers About Obama". *Nytimes.com*. <<https://www.nytimes.com/2008/10/13/us/politics/13martin.html>> (12 November 2019).
- Stockwell, Peter. 2000. *The Poetics of Science Fiction*. London: Longman.
<https://doi.org/10.4324/9781315838878>
- Wiktionary.org. *Wiktionary*. <<https://www.wiktionary.org/>> (7 November 2019).
- Younge, Gary. 2018. "Despite all the warnings, we are normalising Donald Trump". *The Guardian*. <<https://www.theguardian.com/commentisfree/2018/jun/08/donald-trump-president-visits-uk>> (7 November 2019).

From burden to threat

A diachronic study of language ideology and migrant representation in the British press

Gavin Brookes and David Wright

Lancaster University / Nottingham Trent University

This chapter focuses on the ways in which non-native English speakers living in Britain are represented in the British press, and in particular on how these representations have changed between 2005 and 2017. Using a corpus-assisted approach to Critical Discourse Analysis, collocation patterns of the phrase *speak English* reveal that migrants are represented in different ways across the 13-year period, which sees the levels of blame, threat and exclusion levelled at migrants increase and change shape over the years. This chapter builds on previous work by the authors, and emphasizes the importance of re-visiting and adding to corpora when analyzing dynamic discourses, and identifies two different ways in which change can manifest in collocation analysis: through the identification of occasional 'seasonal' collocates, and via consistent collocates being part of different representational patterns.

1. Introduction and background

In April 2018, a 24-year-old woman was attacked on a London Underground train for speaking Spanish. Her two assailants were heard shouting that she should “speak English in England” (Stickings 2018). The incident was described by the British Transport Police as a “vicious and racially-motivated assault” (ibid.). In a separate incident in October that year, a 56-year-old man was arrested on suspicion of racially aggravated actual bodily harm after he punched a woman on a London train because she was speaking Spanish, shouting “you need to speak in English, you’re in fucking England. You shouldn’t speak other languages” (Forrest 2018).

These and other instances of language-based discrimination occur against a UK political and media backdrop which has increasingly targeted people living in Britain who do not speak English as a first language. In terms of the political context, successive UK Governments have implemented a series of policies designed

to ‘encourage’ migrants to the country to learn English to ease their integration (cf. Sebba 2017). This has included such interventions as reductions to state-funded translation of public information materials, the implementation of legislation which places more stringent obligations on migrants to learn (and demonstrate knowledge of) the English language, and suggestions that a deadline should be set by which point everyone living in the country should be able to speak English. From a media perspective, Wright and Brookes (2019) examined the fallout in the right-leaning British national press following the publication of the results of the 2011 Census, which indicated that 8% of UK residents (4.2 million people) had a first language other than English. We found a persistent media narrative which represents non-native English speakers living in the UK in a negative and stigmatizing way, particularly as a burden on and threat towards UK society, ultimately contributing to a wider anti-immigration agenda which legitimizes exclusionary and discriminatory practices against people from minority linguistic and ethnic backgrounds. In this chapter, we take a different perspective on this coverage, adopting a diachronic approach in order to study changes in the right-leaning press discourse surrounding the phrase “speak[ing] English” over a 13-year period (2005 to 2017, with the 2011 Census as a mid-point). Although our focus is on the press within this time frame, we consider the patterns in representation to be symptomatic and reflective of the wider socio-political context, and not just the media and Census alone.

There is a now well-established link between language and notions of national identity and belonging. Piller (2001: 273), for example, writes, “ideologies of national identity are a central facet of modern social identities and they are intricately bound up with linguistic identity”, while Billig (1995: 29) argues that “language is a prime determinant of nationalist identity”. It therefore follows that many individuals “hold passionate beliefs about the importance and significance of a particular language to their sense of national identity” (Blackledge & Creese 2009: 458). Indeed, such beliefs can become deeply passionate and potentially discriminatory, such that language ideologies can have “direct implications for linguistic minorities, their rights, and the policies protecting them” (Vessey 2016: 21). Indeed, the two stories cited at the beginning of this chapter attest to the sorts of behaviours that can manifest at the interface of nationalistic and language ideologies. Given the vested public interest in language-related issues, it is not surprising that they are newsworthy, and it has been argued that the media exploits this newsworthiness and makes “linguistic and cultural diversity visible” (Kelly-Holmes & Milani 2011: 477). To that end, an examination of the media can help reveal the ways in which representations of languages and speakers can operate to discriminate against linguistic minority groups (Blackledge 2002: 84).

In this chapter, we examine press representations of people living in the UK who do not speak English as a first language. Focusing in particular on the right-leaning press, we identify where and how these representations have changed, but also the

ways in which they remain stable, over the period 2005 to 2017. We interpret these representations in terms of the language ideologies they support and propagate. Following Vessey (2017: 278), we understand language ideologies to be

beliefs about languages (or a particular language) that are shared and that become so well-established that their origin is often forgotten by speakers; the beliefs accordingly become naturalized, perceived as common sense, and are socially reproduced.

Moreover, we also interpret these representations and their attendant ideologies in terms of their relation to contemporaneous social and political events, paying particular attention to those which have direct implications for immigration in the UK. Such contextual factors range from the 2011 Census itself, to legislative introductions and the UK Home Office's 'hostile environment' policy towards immigration (cf. Hill 2017).

2. Data and approach

The study reported in this chapter is based on a purpose-built corpus of right-leaning UK national newspaper articles about the topic of speaking English, published between 2005 and 2017 (inclusive). This date range was chosen because it represents the press coverage surrounding the "speak English" debate in the six years prior to and following the 2011 Census. For this chapter, we decided to focus on right-leaning press articles only. This decision was based on two considerations: (i) the right-leaning press accounts for the majority of newspaper discourse in the UK (of the ten national newspapers, six are right-leaning, three are left-leaning and one is centrist; Baker, Gabrielatos & McEnery 2013), and (ii) between 2005 and 2017, the right-leaning press published approximately twice as many articles concerned with the topic of speaking English, as we have defined it (see below), compared with the left-leaning press (4,879 vs. 2,436 articles, respectively).

Articles were obtained from the online newspaper database *Nexis*¹ and were included in the corpus if they contained the words *speak** and *English* (contiguously or non-contiguously) in their headline and/or lead paragraph.² The search

1. Available online at <<http://www.lexisnexis.com/uk>>.

2. The asterisk following *speak* acts as a wildcard for any set of characters, so the search included words like *speaks*, *speaker*, *speakers* and *speaking*. We decided not to include *spoke* or *spoken* in our search term as only a minority of the results produced by these terms (less than 5%) were about speaking English. Instead, these past tense forms tended to be used to frame a quote or to indicate that someone had spoken to the newspaper or the media more widely about the topic of the story.

included all six right-leaning UK national tabloids and broadsheets; *Express*, *Mail*, *Sun*, *Star*, *Telegraph* and *Times*.³ Newswires were excluded, and similar results were grouped together to minimize the inclusion of duplicates (default search settings in *Nexis*). The completed corpus amounts to 4,879 articles (3,044,315 words). Table 1 and Figure 1 provide breakdowns of the number of articles published in each year covered by the corpus.

Table 1. Breakdown of number of articles and words per year

Year	Articles		Words	
	Freq.	%	Freq.	%
2005	233	4.78	145,217	4.77
2006	245	5.02	146,806	4.82
2007	335	6.87	182,504	5.99
2008	316	6.48	176,154	5.79
2009	281	5.76	172,350	5.66
2010	332	6.80	186,221	6.12
2011	394	8.08	225,129	7.40
2012	428	8.77	266,265	8.75
2013	471	9.65	271,721	8.93
2014	550	11.27	355,793	11.69
2015	450	9.22	321,519	10.56
2016	458	9.39	349,139	11.47
2017	386	7.91	245,497	8.06
Total	4,879	100.00	3,044,315	100.00

The number of articles concerned with the speak English debate – as we have defined it – generally increases over time, peaking in 2014 (the year following the publication of the Census results) before falling to levels comparable to 2011 – the year of the Census itself. The corpus includes traditional news articles (i.e. ‘hard news’) but also columns, editorials, reviews, letters to the editor and opinion pieces. We decided to include these publication types since they all occupy space on the pages and websites of the media outlets under study, are all consumed by readers, and all contribute to the language ideologies that we are interested in analyzing in this chapter. We did not filter the results for articles that we deemed not to be

3. Note: Sunday editions and sister publications are subsumed under the main newspaper name, for example *Times* also covers the *Sunday Times*. This is also the case for online versions, for example the *Mail* includes articles published on the website dailymail.co.uk. The exceptions to this are the *Sun* and the *Times*, for which paywalls preclude the inclusion of online articles in *Nexis* results.

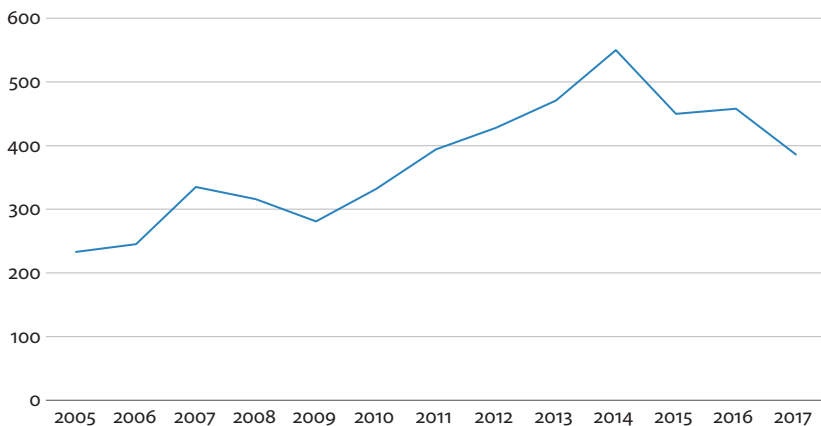


Figure 1. Number of articles about the “speak English” debate published 2005–2017

sufficiently ‘about’ the topic of speaking English, as doing so would have required us to make a series of subjective judgements which would impinge upon the replicability of our data collection procedure (Baker et al. 2013).⁴ Data downloaded from *Nexis* includes a large amount of metadata and indexing information which we removed, leaving only the headline and main body of the articles for analysis.

We analyze this data using a corpus-assisted approach to Critical Discourse Analysis (CDA). CDA is a type of discourse analytical research concerned with the ways in which social power abuse, dominance and inequality are “enacted, reproduced, and resisted by text and talk in the social and political context” (van Dijk 2015: 466; see also Fairclough 1989). The combination of corpus linguistics and CDA is mutually enforcing, with each method able to overcome some of the obstacles associated with the other (Baker et al. 2008). With corpus assistance, CDA can deal more effectively with larger and more representative datasets that are likely to reveal a wider range of representations and account for longer time periods than is possible (or at least practical) using a purely qualitative approach to CDA. Corpus assistance also adds a degree of objectivity to CDA approaches, as it offers a set of predictable analytical techniques and advocates a spirit of methodological transparency underpinned by two guiding principles: (i) no systematic bias in the selection of texts included in the corpus (i.e. do not exclude a text because it does

4. This point notwithstanding, we performed a ‘sanity check’ on the corpus to ensure that the proportion of false positives (i.e. articles that clearly had little to do with the English language) would not unduly influence our analysis. To do this, we extracted a random sample of 100 articles spanning all years and publication types. From this, we determined that just 3 articles had limited relevance to our purposes but the remaining 97 had something to do with the English language and those who can or cannot speak it (well).

not fit a pre-existing argument or theory) and (ii) total accountability (all data gathered must be accounted for) (McEnery & Hardie 2012).

At the same time, our corpus-assisted approach also stands to benefit from its synthesis with CDA, and particularly relevant here is CDA’s commitment to analyzing social context (van Dijk 2001), which includes providing arguably more robust theoretical accounts of the ways in which texts interact with the contexts in which they are produced and consumed. In the present study, we draw upon our knowledge of the social and political backdrop against which the articles in our corpus were published, paying particular attention to UK language policy and relevant events taking place within the time span of our data, in an attempt to re-situate the news stories within the original contexts in which they were produced and consumed.

Our analysis takes a collocation-based approach to trace similarities and changes in the discourse surrounding the key expression *speak English* ($n = 3,398$) between 2005 and 2017. The frequency of this phrase in each year of the corpus is provided in Table 2.

Table 2. Frequency of *speak English* in each year of the corpus

Year	Freq.	%
2005	110	3.24
2006	124	3.65
2007	205	6.03
2008	160	4.71
2009	186	5.47
2010	197	5.80
2011	237	6.97
2012	264	7.77
2013	363	10.68
2014	453	13.33
2015	408	12.01
2016	323	9.51
2017	368	10.83
Total	3,398	100.00

We decided to use the phrase *speak English* as an analytical entry point as it corresponds with the wording of one of the 2011 Census items (i.e. “How well do you *speak English*?”), while also reflecting the search term used to source articles for the corpus. Using *WordSmith Tools version 7* (Scott 2016), we generated a list of collocates of the phrase *speak English* ($L5 > R5$) and extracted collocational pairings with an MI³ score of at least 11 (Brezina, McEnery & Wattam 2015). This measure

was considered preferable to frequency alone as it promotes lower-frequency but more exclusive associations which can benefit the identification of representations associated with a word, while curbing the low frequency bias associated with straightforward MI. Therefore, no matter how suggestive, an association which is not repeated enough will be less influential than an association that is more firmly established in the corpus. Furthermore, MI³ avoids the sensitivity to corpus size and calculation of significance scaled to a particular range of values associated with t-scores and log-likelihood, respectively (see Brezina et al. 2015 and Gablasova, Brezina & McEnery 2017 for comparisons of measures). We then grouped these collocates into themes based on manually reading a random sample of 100 uses per collocate (where possible) in which each occurred alongside the expression *speak English*. Based on this analytical step, we grouped the collocates into the following thematic categories: (i) proficiency, (ii) multilingualism, (iii) learning English and integration, and (iv) public services and the private sector. We monitored change over time by identifying which years any given collocate did or did not reach our MI³ cut-off value of 11. This allowed us to observe collocates which had consistent prominence across the time frame as well as those which appeared only at given points in the data, and perhaps reflect growing or waning trends in the reporting. However, we take a broad view of change that extends beyond quantitative changes over time. In our analysis, change manifests in two ways. One way is the presence and absence of particular collocates at particular points in time, with some collocates meeting the threshold of 11 in only one or two years of the corpus. Meanwhile, the other acknowledges that although the same collocate may appear across many years, the *way* in which it is used can change, in terms of the different representational patterns it contributes towards. This diachronic approach to corpus-aided discourse analysis differs from existing methods, most obviously Gabrielatos and Baker's (2008) analysis of 'consistent collocates' which appear in at least seven of their ten annual sub-corpora of newspaper articles about refugees and asylum seekers. This filtered out so-called 'seasonal collocates' which were frequent in some years and not others. Such collocates were undesirable for Gabrielatos and Baker's study, but form an important part of our analysis here. On a further methodological note, our findings also highlight the importance of returning to and expanding upon previously analyzed corpora (in this case from Wright & Brookes 2019) which, as Baker and McEnery (2019) argue, can be useful for monitoring changes in a dynamic discourse (in our case, around the *speak English* debate), and to assess the ways in which media representations reflect but also potentially influence their socio-political contexts.

3. Analysis

In this section, we examine each of the four themes identified through our grouping of collocates: proficiency; multilingualism; learning English and integration; and public services and the private sector. The analysis is structured thematically and addresses each of these themes in turn.

3.1 Proficiency

The first set of collocates we want to consider consists of terms relating to the proficiency with which migrants are represented as being able to speak English (see Table 3). This constitutes a large number of grammatically diverse collocates, including modal verbs (*can, cannot, can't, could, couldn't, must, should*), auxiliary verbs (*did, didn't, do, does, doesn't, don't*), degree adverbs (*barely, fluently, properly, well*), lexical verbs (*struggle, struggled*), adjectives (*able, unable*) and a noun (*standard*).

The first thing to note about this table is the proliferation of negative modal and auxiliary verbs. *Can't, cannot, couldn't, didn't, doesn't* and *don't* all collocate with *speak English* to refer to people who do not speak English, while the affirmative forms *can, could, do*, and *does* tend overwhelmingly to be negated to similar effect (e.g. “those who *do not* speak English as a first language” (*Times*, 2007)). In a similar vein, the adjective collocate *unable* is consistently used to describe migrants as being unable to speak English, with the affirmative *able* constantly negated (e.g. “ONE in 10 doctors in Britain may not be *able* to speak English properly” (*Star*, 2007)). We note here that this is a fairly stable feature of the corpus, with most of these terms emerging as strong collocates of *speak English* in the majority of the years, while *can't* and *don't* appear as strong collocates across all years.

That these articles focus on people – migrants specifically – who do not or are not able to speak English is not particularly surprising, as it is not only arguably more newsworthy in a cultural context where English is the dominant language, but also reflects the tendency of the press – and the mainstream media generally – to report stories about things and events that might be considered (or which it can frame as) problematic or in some way negative (Bourdieu 1998). Yet taken together, these collocates also reflect what we have previously described as a “deficit model of linguistic competence” (Wright & Brookes 2019: 67), which focuses on migrants’ linguistic deficiencies at the expense of their capabilities. In other words, rather than describing migrants as speakers of English as an additional language or indeed as speakers of other languages, collocates such as *can't, cannot, unable*, and so on, help to foreground migrants’ perceived inability to speak English, all the while backgrounding or obscuring details about the languages they *do* speak.

Table 3. Proficiency collocates of *speak English* ($MI^3 \geq 11$) with normalized frequency of collocation (per hundred occurrences of *speak English*)*

Collocate	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
<i>able</i>			4.87			4.06	5.90	6.06	2.20	3.53	4.90	6.19	8.15
<i>barely</i>	4.54				3.76	2.53	2.53	2.27	1.37		1.83	0.80	1.34
<i>can</i>	11.82		4.87	7.50	6.45	10.65	5.06	10.60	3.86	3.75	7.35	4.33	7.07
<i>cannot</i>	10.00	7.25	7.80	8.75	10.21	7.10	13.50	3.03	9.64	9.27	10.34	17.33	
<i>can't</i>	7.27	9.67	8.78	6.25	5.91	4.56	6.32	4.16	6.06	6.84	3.09	1.21	9.63
<i>could</i>	10.00	8.06	4.39		5.91		4.22	7.95	10.19	6.84	2.69	6.50	6.25
<i>couldn't</i>				3.75	4.30			3.40		1.32	2.26	5.77	0.83
<i>did</i>									2.48		3.19	4.02	3.80
<i>didn't</i>									1.93	1.99	2.21		16.86
<i>do</i>		6.45	14.63	14.37	8.60	12.18	8.44	6.06	11.85	9.49	7.84	9.29	2.43
<i>does</i>					6.45		3.79			2.21	2.70		2.72
<i>doesn't</i>								3.41		3.53	4.66	2.48	2.72
<i>don't</i>	5.45	6.45	5.85	10.00	6.45	6.09	5.91	3.79	9.36	6.84	16.34	2.94	9.51
<i>fluently</i>									2.20		0.30		
<i>must</i>						5.58	3.37		3.58	5.51	6.13	5.77	
<i>properly</i>		5.64				6.09	3.79	3.78	3.03				2.87
<i>should</i>										4.19	0.44	4.33	
<i>standard</i>											15.89		
<i>struggle</i>							2.11			1.32			
<i>struggled</i>													1.73
<i>unable</i>									2.20	1.98	0.69	2.24	6.33
<i>well</i>						4.06			8.81	4.86	2.70		

* Blank cells indicate that a collocational pairing either did not occur in that year of the corpus or was assigned an MI^3 score below 11, meaning that it did not meet our threshold for significance.

Further analysis reveals an interesting trend with regard to *who* is represented as not speaking English at different points across the timespan represented by the corpus. Taking *can't* as a case in point (given its consistent collocation with *speak English* over time), an analysis of the social actors and groups to whom this refers shows that in each of the first four years of the corpus (2005, 2006, 2007, and 2008), it is foreign sportspeople working in the UK who are represented as not being able to speak English more than any other group. This consisted mainly of football players, but also football coaches and cricketers. In the next two years of the corpus (2009 and 2010), schoolchildren – children of migrants specifically – are most frequently described as not being able to speak English following the publication of Government statistics about the languages spoken by schoolchildren in the UK. Then in 2011 and 2012, it is healthcare staff who “cannot” speak English most often. In the next year, 2013 – the year of the Census results publication – the focus

moves away from specific individuals and groups and it is “migrants” generally who are most frequently represented as not being able to speak English. Yet it was not just *who* was the focus of this type of reporting but also *how* they were represented that changed over time, with migrants written about in more general terms in the years following 2013, for example as “people living here” (see (1)). Moreover, as Example (1) also demonstrates, at this point the articles began to provide more precise quantification of the number of migrants who could not speak English – figures which not only reflect the Census outcome but also add authority and legitimacy (Potter, Wetherell & Chitty 1991) to claims that the English language is in “decline” in the UK. Note that examples provided in this chapter were selected because they were deemed to be representative of the wider patterns being discussed.

- (1) Polish is the second language of Britain and nearly 140,000 people living here *can't* speak English at all (Mail, 2013)

However, migrants were not just represented as not being able to speak English at all but were also construed as having *limited* English skills – that is, as being able to speak English but only to a certain standard. Returning to Table 3, we find evidence for this in the shape of the lexical verbs *struggle* and *struggled*, the noun *standard* and the degree adverb *barely*, as well as *fluently*, *properly*, and *well*, which tended to be negated.

- (2) The census also showed that nearly a third of native Punjabi and Bengali speakers also could not speak English fluently. Highlighting the additional strain placed on the NHS, a third of migrants unable to speak English *fluently* were in poor health (Mail, 2013)
- (3) Lecturers ‘*struggle* to speak English’ at elite universities charging students £9,000 a year (Mail, 2014)
- (4) The new law, revealed in the Immigration Bill yesterday, prompted critics to warn it showed the extent to which immigration from Europe was out of control. Tory MP Andrew Percy said: ‘People should not be allowed to come into this country to work at all unless they can speak English *to the right standard*.’ (Mail, 2015)

As these examples show, from the perspective of some of the articles in our corpus, it is not enough for migrants to speak English, but they are also expected to speak English *fluently*, to a particular *standard* and to not *struggle*, else they are liable to be targeted and blamed for issues such as placing a strain on the NHS and harming university education (we return to the issue of public services later in the analysis). However, the discrimination against national varieties of English reported in the United Kingdom and other contexts (Lippi-Green 2012), indicates that the notion

of fluency is contested. In other words, what counts as “fluent” or, to quote Andrew Percy in Example (4), the “right standard” of English, can be debated. However, the failure by these and other articles which represent migrants as *struggling* to speak English or as not speaking it *fluently* enough or to a high enough *standard* to describe how they perceive or measure such notions, or to acknowledge the contested and complex nature of such debates, leads to an oversimplified representation of fluency. This, in turn, has the potential to widen possibilities for linguistic discrimination, since potentially anyone who speaks English non-natively and “with an accent” (Lippi-Green 2012) is liable to be perceived as non-fluent and so to be blamed for the kinds of issues raised in Examples (2) to (4).

Having established migrants not speaking English – at least to a high standard – as problematic, articles published in the latter years of our corpus also adopt more of a prescriptive tone, as indicated by the deontic modal verb collocates which are used to assert that migrants *must* and *should* speak English (well), as in these examples of *must speak English*:

- (5) Labour pledge to pass Calman by next spring; Public workers *must* speak English, says Brown (Times, 2010)
- (6) Boris: Migrants who want to enter Britain *must* speak English (Telegraph, 2016)

The earliest year in which any of these terms emerge as strong collocates of *speak English* is 2010 (*must*); the year preceding the Census. This suggests that the press’s more prescriptive tone emerged – or at least intensified – in the run-up to the 2011 Census, at which point the articles began to foreground the idea that migrants *ought to* be speaking English and to a high standard. *Must* also emerges as a strong collocate in 2011 but not in 2012. It then re-emerges in 2013 – the year in which the Census results were published – and remains a strong collocate for the three years following (2014, 2015, and 2016), during which time the prescriptive *should* also remains a strong collocate. The prescriptive notion that migrants *should* and even *must* learn to speak English to enter and live in the UK could therefore be detected before the census. However, the uninterrupted strong association between these collocates and *speak English* in the years immediately following the publication of its results might indicate that this prescriptive attitude was exacerbated or even legitimated by the Census outcome.

Taken together, then, the collocates analysed in this section suggest that for the duration of the time period represented by our corpus, the right-leaning press framed migrants’ proficiency from a perspective of deficiency which foregrounded their linguistic limitations while backgrounding, if not obscuring altogether, their capabilities. While this was a fairly constant feature of the press discourse

across this period, our analysis has also revealed shifts in the representations over the years. Namely, the focus on the type of migrants who cannot speak English changed over time – from individual and exceptional cases like sportspeople in the public eye, to the broader groups of schoolchildren and healthcare staff and then, following the publication of the Census results in 2013, to the broadest group – all migrants living in the UK. Another feature of the articles published after 2013 was the problematization of migrants not being able to speak English with ease and fluency, and the adoption of a prescriptive tone to foreground the view of the press, and the various public figures they cite, that migrants *must* speak English, and they must speak it well.

3.2 Multilingualism

Although we have observed that some collocates of *speak English* foreground a deficiency model of linguistic competence, other collocates do suggest a more nuanced take on this issue. This next set of collocates we want to explore consists of terms which denote whether English is spoken as a first or additional language. This category of collocates – which broadly surround the theme of multilingualism – includes the words *additional*, *first*, *language*, *mother*, *native*, *second*, and *tongue* (Table 4).

Examining the texts in which these terms collocated with *speak English* across all the years of our data, it emerged that all of the collocates in Table 4 exhibit a similar pattern; that is, they tended to occur in passages reporting statistics relating to the number of schoolchildren living in the UK who speak English as a second or additional language. Taking these terms together, this pattern accounted for 87% of the articles in which one or more of these terms collocated with the phrase *speak English*.

- (7) In the last 12 months, the number of children who speak English as an *additional language* has risen by almost 54,000, according to statistics published by the Department for Education. The National Union of Teachers has called for funding to help support the teaching of these children to be protected from Government cuts. (*Express*, 2013)
- (8) The Department for Children, Schools and Families statistics showed that 12.5 per cent did not speak English as their *mother tongue* as of January this year. For primary schools, the proportion was 14.3 per cent, compared with 10.5 per cent in 2004, the year before European Union expansion. (*Mail*, 2008)
- (9) A CITY'S primary schools face "meltdown" as nearly half its children do not speak English as a *first language*. In Bradford, 43% of pupils' primary language is not English, says the Department of Education. And in its secondary schools the total has soared to 30.3%. Pam Milner, of the Bradford branch of teachers'

Table 4. Multilingualism collocates of *speak English* ($MI^3 \geq 11$) with normalized frequency of collocation (per hundred occurrences of *speak English*)

Collocater	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
<i>additional</i>									1.37	1.32			
<i>first</i>			7.80	5.63	13.44	16.24	6.33	8.71	7.99	6.40	3.68	4.33	10.05
<i>language</i>		6.45	11.22	11.25	13.44	24.63	8.44	15.90	19.00	20.10	4.41	8.99	11.68
<i>mother</i>				4.37									
<i>native</i>								1.89					
<i>second</i>		4.83		3.75		7.61		4.54	4.40	2.69		3.10	2.99
<i>tongue</i>				5						0.73			

union NASUWT, said ministers should come to the city to see its problems first hand. It also needed to be protected from £6billion public spending cuts proposed by the new coalition government. (Star, 2010)

- (10) ENGLISH IS FOREIGN TO 2 IN 3 KIDS Extra tuition costs a fortune. The immigration explosion is crippling British schools as staff struggle to teach children who cannot speak English. A shocking 30% of pupils in Manchester now speak English as a *second* language, and that figure rises to two in three in some parts of the city. (Star, 2010)

The articles from which these extracts were taken follow a series of Government reports regarding the number of UK schoolchildren who spoke English as a second or additional language at various points across the time span of our corpus. As these examples show, the majority of these articles (64%) also referenced the economic costs that these children were purported to entail for schools and – by extension – taxpayers. For example, in (10) the reported “extra tuition” required by these children is described as costing “a fortune”. The volume of children with English as a second language, and the (economic) cost that this is purported to entail for schools and taxpayers, was also framed in explicitly negative evaluative terms, in the above as causing “problems” and “meltdown” (9), as well as “shocking” and “crippling” schools (10).

This economic perspective was not time-specific but could be found across all years featuring the collocates in Table 4. However, one aspect of the representation which did change over time was the solutions to these concerns that were put forward by the articles. Prior to 2014, as shown in (7)–(10), the response of the right-leaning press was to call for the Government to protect and even increase funding for schools. In some cases, such appeals were embedded within quotes from public figures involved in the education sector, as in (7) and (9). However, in the 2014 reportage, concerns about school funding seem to be replaced by concerns that children with English as a second or additional language are actually

“outperforming” or “overtaking” children who speak English as their first language, as in (11) and (12):

- (11) AN INNER-CITY state primary, where most of the pupils speak English as a *second language*, is outperforming nearly all the fee-paying preparatory schools in The Sunday Times league tables. (Times, 2014)
- (12) Children who speak English as a *second language* are outperforming native speakers for the first time in GCSE exams, according to latest figures. (Mail, 2014)

We observed this representation in 36% of the articles published in 2014 featuring one or more of the collocates in Table 4, compared to just 5% of the articles published prior to this year. Part of the reason for this was the publication of Government figures in 2014 which showed that primary school children with more than one language had, for the first time, achieved higher grades for some subjects compared to children who spoke only one language. It is possible, then, that the publication of these figures helped to prompt change in the press discourse around this time. Where migrant children were previously construed as struggling and so burdening schools’ resources, following this report they are more likely to be depicted as competing and with and even out-performing children with English as a first language. We interpret this change in the representation of schoolchildren as reflecting a continuation of the longer-term division between children based on their first language but by this time this division is framed in terms of threat, rather than as a reason to protect public spending.

3.3 Learning English and integration

The verb *learn* is a pervasive collocate of *speak English* throughout the corpus, meeting or exceeding the threshold MI³ score of 11 in eight of its 13 years, appearing with such a strong association for the last time in 2015 (Table 5). Although the

Table 5. Learning collocates of *speak English* (MI³ ≥ 11) with normalized frequency of collocation (per hundred occurrences of *speak English*)

Collocate	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
<i>attend</i>									2.20				
<i>learn</i>	5.45		5.05		4.37		4.80	3.86	6.89	2.46	3.99		
<i>learned</i>											7.30		
<i>teach</i>						3.09							
<i>effort</i>									2.20				
<i>quickly</i>									1.38				

relative frequencies of *learn* as a collocate of *speak English* remain fairly consistent across time, the representational uses to which it is put do not.

In 2005, *learn* appears in the context of *speak English* five times in total, 2% of all articles in that year of the corpus. It appeared in articles about a British TV presenter having elocution lessons so audiences in the US could understand him, Americans not understanding British slang, and a footballer endeavouring to learn English in order to prevent his teammate transferring away from their club. Notably, there is one instance in an article about immigration and integration and in particular plans of the Labour Government at the time to manage migration (see (13)).

- (13) It [Labour] aims to continue to ensure that those who settle in Britain can play a full part by helping them to speak English and *learn* about British life.
(*Times*, 2005)

In this example, although *learn* collocates with *speak English*, it relates to learning about *British life*, rather than the English language. The article is about plans to help migrants speak English, reflecting a responsible, active role of the Government in that process. However, this supportive perspective does not persist over time, and in the 2007 data the collocation of *learn* with *speak English* is used to shift the responsibility for learning English and integrating into society away from the Government and onto the migrants themselves, as in (14).

- (14) GORDON Brown has warned immigrants that if they want to stay in Britain they will have to *learn* to speak English. The Prime Minister made it clear that a condition for being part of British society must be a willingness to take on certain responsibilities. And those should include speaking English and understanding British cultural traditions.
(*Star*, 2007)

This is the first point in the data at which non-native English-speaking migrants learning English is linked to their “being a part of British society”, with this language learning and assimilation framed as being *their* responsibility. This stance differs considerably to that in the 2005 example (13) and is the launching point for a long and fluctuating relationship in the (right-leaning) media between migrants’ responsibility for learning English and their integration in British society, as in (15) to (17).

- (15) Migrants lingo call – MORE immigrants should *learn* to speak English to improve community relations, says Eric Pickles.
(*Sun*, 2012)
- (16) IMMIGRANTS must *learn* to speak English properly to stop communities breaking down, a Cabinet Minister said yesterday. Westminster Communities Secretary Eric Pickles warned that newcomers with “broken English” could not play a full role in society.
(*Express*, 2013)

- (17) The Prime Minister criticised segregation between communities, promising new steps to promote integration in schools and housing estates and to ensure minorities *learn* to speak English. (Telegraph, 2015)

In (15) from 2012, Eric Pickles – the then-UK Secretary of State for Communities and Local Government – is represented as imploring migrants to learn English. At this point in time, it was suggested that migrants should learn to speak English to “improve community relations”, positioning migrants’ ability to speak English as potentially improving a situation which is not (explicitly) framed as being bad. A year later, Pickles raised the stakes, and migrants’ English language skills were no longer simply a means of improving community relations, but they were now responsible for preventing communities from “breaking down” entirely (16). As such, the sense of seriousness and urgency is increased, and migrants now *must* learn to speak English (see Section 3.1). At the same time, not only are migrants implored to learn English as they were in 2012, but in 2013 they must learn to speak English *properly*, with the introduction of an ill-defined, vague level of proficiency now required. As shown in (17), by 2015 these sentiments were shared by then-Prime Minister David Cameron, who suggested that the breakdown of communities prophesized by Pickles had by now been realized, as he “criticised segregation between communities”. A response to this segregation, according to Cameron, was the promotion of integration, which included ensuring “minorities learn to speak English”. Thus, the responsibility of migrants and their English language abilities had shifted once more. No longer were they the last line of defence against communities breaking down; they now held the solution to repair already broken communities.

Discussions around speaking English and integration did not occur in a vacuum. Between 2007 and 2011 they were accompanied by right-leaning media reports propagating the idea of non-native English speakers having access to benefits withheld by both Labour and Conservative-led Governments:

- (18) ‘Learn English or lose your benefits’ – ETHNIC minorities could face losing their benefits if they do not *learn* to speak English, ministers said yesterday. Welfare Minister Jim Murphy warned that up to 40,000 could be struggling to find work because they are unable to master English. (Mail, 2007)
- (19) UNEMPLOYED people who cannot speak English will be forced to *learn* the language or risk losing their benefits, David Cameron said last week. (Telegraph, 2011)

In the 2007 data, this appears in articles reporting Labour welfare proposals to restrict benefit payments to migrants on the basis of their proficiency in English (18), and in 2011 in reports of similar Conservative plans (19). Such plans are based

on a framework that sees non-native English speakers as an economic burden on the UK, the logical extension being that they are undeserving of taxpayer-funded welfare support, precisely on the basis of their perceived lack of English proficiency. By focusing on unemployed people, reports are inherently limited to a small sub-section of non-native English-speaking migrants, and although their being unemployed makes them an easy target for vilification in the media (Baker & McEnery 2015), the generalizability of this vilification is limited. Although such welfare plans are reported on intermittently in later years (in 2015 there are two articles about the Conservative Government's "incentives and penalties" for forcing migrants to learn English which included restricting housing benefit), in 2012 and 2013 they give way to a focus on the cost of translation services:

- (20) Emma Boon, campaign director of the TaxPayers' Alliance, said: "Taxpayers will be shocked that so much is being spent on translation and interpretation in the NHS. "They expect their money to be going towards treatment for sick people, not on language services. [...] "But those who live in Britain should make an effort to *learn* to speak English so that they are not burdening services like the NHS with ongoing costs for translation." (*Mail*, 2012)
- (21) "Those needing translation, in many cases, have lived in this country for many years and therefore have a responsibility to learn English". Matthew Elliott, chief executive of the TaxPayers' Alliance, said: "Those who choose to live in Britain must make an effort to *learn* to speak English so that council taxpayers do not get burdened with these costs". (*Express*, 2012)

By shifting the source of the financial burden from unemployment benefit to translation services, the net of responsibility is widened, and more migrants are made responsible for costing taxpayers' money. The fact that the new target is translation services specifically should not be overlooked. Examples (20) and (21) suggest that translation services should be reduced and angering and frustrating "the taxpayer" is a means by which to get public support for this idea. Sebba (2017) notes that the reduction in translation services has been on the UK Government agenda since as early as 2007, as the provision of translated materials was considered a barrier to migrants learning English and therefore an obstacle to integration and something that encouraged segregation. However, the argument for stripping publicly funded translation services is at odds with the concurrently unfolding claims that migrants' lack of English language proficiency is damaging integration. While it may be that a political viewpoint holds that translated materials undermine community cohesion, translation and translated materials can be valuable resources in the acquisition of English as a second or additional language (e.g. Cook 2010). Therefore, at the same time as migrants are being incrementally blamed for the breakdown of communities in the pages of the right-leaning press, there are calls for the removal of

resources and materials which could help arrest this perceived problem. The important point here, then, is that in the landscape of the speak English debate between 2005 and 2017, these two seemingly incompatible representations emerged at the same time and on the same side of the political spectrum.

3.4 From public to private services

The final section of our analysis explores collocates relating to the theme of public services and the private sector, shown in Table 6.

In Wright and Brookes (2019), we identified the ways in which the press transported non-native English-speaking migrants into readers' everyday worlds by focusing on social actors in public service contexts, namely schools and healthcare. The most obvious pattern emerging from the collocates in Table 6 is the consistency with which *children* and *pupils* appear across the years. With *pupils*, which has the strongest association with *speak English* of these two collocates, 2006, 2008, 2010, and 2014 are the years with the highest relative frequency. Despite being punctuated by intervals of two years or more, the patterns of representation in which this collocation is found remain consistent and reflect the right-leaning media's focus on the numbers of schoolchildren who have English as an additional language, as exemplified in (22) to (25).

- (22) *PUPILS* who speak English as their first language are in the minority at most inner London primary schools. Nationally, one in five primary pupils is from an ethnic minority, making non-English speakers a majority in many schools. (Mail, 2006)
- (23) ONE in every eight *pupils* in our schools now speak English as their second language, a shock report revealed yesterday. (Star, 2008)
- (24) THE immigration explosion is crippling British schools as staff struggle to teach children who cannot speak English. A shocking 30% of *pupils* in Manchester now speak English as a second language, and that figure rises to two in three in some parts of the city. (Star, 2010)
- (25) According to figures, 1.12m children now speak English as a second language, including 654,405 in primary schools, 455,205 in secondaries and 13,585 in special schools. (Telegraph, 2014)

This represents an unchanging media narrative, with a pre-occupation with the "rising" numbers of non-native English schoolchildren in schools and the negative effect of this rise (see earlier analysis on school funding in Section 3.2). Although in the public sector context of schools the reporting mainly focuses on the language of *pupils* rather than teachers, there has been sustained right-leaning media attention on the first language of NHS *doctors*, as in (26) and (27).

Table 6. Services collocates of *speak English* ($MI^3 \geq 11$) with normalized frequency (per 100 occurrences of *speak English*)

Collocate	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
<i>banned</i>					6.01		3.49						
<i>banning</i>					3.28								
<i>changing</i>										1.12			
<i>character</i>										1.12			
<i>children</i>	6.36			6.54	6.56	14.43		9.27	4.68	8.72	2.99	2.8	4.1
<i>crime</i>													3.01
<i>customers</i>					14.21								
<i>dealing</i>											1.5		
<i>doctors</i>							10.04		1.65	1.57			
<i>drivers</i>													2.46
<i>gps</i>						2.58							
<i>jail</i>										1.12			
<i>kids</i>						3.09						1.56	
<i>nurses</i>	4.55						6.11				2		3.01
<i>patients</i>							3.93						2.73
<i>primary</i>					3.83	4.12							
<i>priority</i>													1.37
<i>pupils</i>	10.66	6.57	10.46	9.29	13.4	3.49	4.25	5.23	10.07	2.24	2.8	2.73	
<i>required</i>											2.24		
<i>roles</i>											1.75		
<i>schoolchildren</i>						2.58		2.2					1.37
<i>serve</i>					4.92								
<i>services</i>											2		
<i>victims</i>													1.91
<i>waiters</i>													1.64
<i>wife</i>										2.01			
<i>witnesses</i>							2.62						
<i>workers</i>						5.15				1.57	2.99		

- (26) FOREIGN *doctors* who cannot speak English are to be banned from working in NHS hospitals and clinics, the Health Secretary announced.

(*Telegraph*, 2011)

- (27) NHS *doctors* who can't speak English face being struck off. HEALTH staff who cannot speak English properly face being struck off from the NHS under plans revealed today.

(*Express*, 2014)

The noun collocate *doctors* has a particularly strong association with *speak English* in 2011, 2013, and 2014 (Table 6). In all three of these years, almost all of the co-occurrences of *doctors* with *speak English* relate to Coalition Government plans

to ban doctors who cannot speak English from practising in the NHS, as shown in (26) and (27). *Nurses* working in the NHS are subjected to similar treatment (see (28) to (30)).

- (28) FOREIGN *nurses* who cannot speak English properly are putting patients' lives at risk, a coroner warned yesterday. (*Mail*, 2005)
- (29) In the latest lunacy issued by the EU, doctors and *nurses* who struggle to speak English are being allowed to work in our hospitals. (*Express*, 2011)
- (30) NIGEL FARAGE has called for NHS doctors and *nurses* who do not speak English to be sacked. (*Telegraph*, 2015)

As is represented by these examples, the attention shifts between the risks posed by non-native English-speaking nurses, to questioning why they are still working in UK hospitals, to explicit calls for their “sack[ing]”. Such representations preceded, and perhaps contributed towards, policy and legislative change in 2016, as Part 7 of the Immigration Act of 2016 introduced specific English language requirements for migrants working in public sector roles, stipulating that: “a public authority must ensure that each person who works for the public authority in a customer-facing role speaks fluent English”. Interestingly, after the introduction of this law, the focus on nurses changes. The collocation of *nurses* with *speak English* is not a strong association in the 2016 data (at least by our criteria) but returns in 2017, by which point this pairing is used not in relation to NHS workers, but almost exclusively in articles covering a story from a care home (see Example (31)).

- (31) A CARE home where Polish-born *nurses* could not speak English properly to patients, with one having to use a translation service, has been criticised by inspectors. (*Express*, 2017)

Extract (31) is taken from a story about Polish nurses working in Penrhos Polish Care Home in Gwynedd (Wales) in which, according to reports, inspectors found that two of the eleven nurses “struggled to speak English to residents” (*Express*, 2013). Notwithstanding that this is a *Polish* care home, and that the nurses are judged to have not been able to speak English “properly” as measured against ill-defined standards of what counts as “good enough”, this story is particularly noteworthy because this care home is privately owned by Polish Housing Society Limited, a registered charitable organization. While these nurses without “fluent” English may have been “banned” from working in an NHS care home, no such legal requirements existed for the private sector at this time. Yet what we observe here is a familiar representation of nurses in the public sector, except that here it is being applied to the private sector, too.

In addition to private healthcare organizations, this representation is also evident in relation to two other customer-facing private sector occupations; namely, *waiters* and (taxi) *drivers*:

- (32) “Co-existence can only work in Germany if we all speak German,” Jens Spahn, seen by many as a potential successor to Angela Merkel, said. [...] “It drives me up the wall the way *waiters* in Berlin restaurants only speak English,” he told Neue Osnabrücker Zeitung newspaper. (*Telegraph*, 2017)
- (33) In July, state laws which remove the requirement for *drivers* to speak English will overrule the local requirement in Miami Dade County. (*Mail*, 2017)

In relation to both waiters and taxi drivers, the *Telegraph* and the *Mail* report on contexts outside of Britain. In (32), the *Telegraph* reports on an interview with Jens Spahn, the Parliamentary Secretary of State for Finance in Germany, in which he laments the use of English by waiters in restaurants in Berlin. All six instances of *speak English* with *waiters* occur in articles covering this story. In another story covered by the *Mail*, as in (33), the focus is on an incident in Miami (U.S.), where a Spanish-speaking Uber driver, Carmen Echevarria, was fined \$250 under local law after a customer complained to Miami airport staff that Echevarria could not speak English. Both these stories are newsworthy to target readers of British right-leaning newspapers not only due to Germany’s and the U.S.’s respective geographical and cultural proximity to the UK (Bell 1991: 156) but also because of the likely “value proximity” (Balmas 2017: 668) of these stories to their readers, wherein “value” refers to “deeply rooted motivations or orientations guiding or explaining certain attitudes, norms, and opinions”. By drawing on stories from elsewhere in the world where private sector workers are chastised for not speaking the main language of their country of residence, such beliefs and opinions about non-native English speakers working in the UK are shared, supported and rationalized. In doing so, the right-leaning media may be laying the groundwork for a new narrative and public debate in which non-native English-speaking private sector workers can be demonized and punished, as they are elsewhere in the world, possibly with the aim of displacing them from their jobs and roles within British society. Indeed, there is evidence of this already, in that the only other story in 2017 in which *speak English* co-occurred with *driver* was concerned with announcements of new English Language requirements for drivers with the taxi firm Uber working in the UK:

- (34) Uber has lost the High Court challenge and its *drivers* will have to speak English. (*Mail*, 2017)

The new requirements, which came into full effect in April 2019, demand that Uber drivers have a UK qualification in English (a GCSE A-G as a minimum), thus

excluding those who do not hold such a qualification from working as a driver for the company.

It is not only public and private sector workers who are reported on, ostracized and castigated in the press for their lack of (native) English – such representations also exist for customers and users of such services. In 2009, our data is characterized by a strong association between *speak English* and *customers* in which all instances of this collocation are found in articles (spanning all newspapers in the corpus) reporting on a postmaster who refused to serve customers at the post office where he worked unless they spoke English:

- (35) AN IMMIGRANT postmaster has BANNED *customers* who cannot speak English – so they learn the language. Sri Lankan-born Deva Kumarasiri said: “If you come to Britain you have got to speak English. It’s as simple as that”.
(*Sun*, 2009)
- (36) I’M BRITISH AND PROUD; Post Office patriot is your hero – SHOPKEEPER Deva Kumarasiri has been hailed a Great Brit hero after banning *customers* who do not speak English.
(*Star*, 2009)
- (37) Sri Lankan-born Deva Kumarasiri, who’s interested in the efficient running of his post office and is an integrated foreigner to boot, has the courage to say what needs to be said – that his *customers* must speak English.
(*Express*, 2009)

This is a useful story for the right-leaning press, since the integration of Kumarasiri’s voice, himself a non-native English-speaking migrant, into the reporting amplifies its newsworthiness (Bednarek 2016) but also legitimizes (van Leeuwen 2007) the argument that all migrants should speak English. In other words, presenting a migrant as holding these opinions allows the right-leaning media to express and justify exclusionary behaviours (such as banning people from shops) through the reporting of speech for which they have no accountability, and allows the creation of a “good” or “model” migrant in Kumarasiri (who can speak English) as distinct from “bad” migrants (who cannot). To this end, the identity of Kumarasiri as “AN IMMIGRANT postmaster”, “Sri Lankan-born” and “an integrated foreigner” all emphasize his non-Britishness, and he is offered as a spokesperson for migrants but one who says “what needs to be said” (37). Kumarasiri’s discriminatory practices (for which he was eventually dismissed from employment) are not only rationalized and justified by the press but are celebrated. For example, in the *Star* he is described as a “Post Office patriot”, “your hero” and “a Great Brit hero” (36). Such representation makes explicit the newspapers’ supportive stance on banning people from shops on the basis of their language abilities, at the same time invoking a nationalistic rhetoric to present Kumarasiri’s stance as patriotic and thereby equate speaking English with Britishness.

The banning of customers from shops is not something that persists in the years following 2009 to the extent where *speak English* and *customers* is a strong collocation. However, what does emerge from the most recent data (2017) is a familiar representation relating to public services. Only this time, it is not just public sector workers who are targeted, but the users of those services, with a particular focus on how non-native English-speaking migrants are supposedly being given preferential treatment by the NHS and the police:

- (38) *Patients* who cannot speak English are given double the appointment time at the GP because it takes so long to translate, doctors have admitted. In recent years GPs' surgeries have struggled to keep up with soaring patient numbers, fuelled by immigration, which has seen an extra two million National Insurance registrations since 2011. Many GP lists are now 'closed' and patients complain that they cannot get appointments because of the influx. (*Telegraph*, 2017)
- (39) A SENIOR police officer sparked anger last night by suggesting *crime victims* who don't speak English could be given the highest *priority*. (*Express*, 2017)
- (40) English in cop snub – *CRIME victims* unable to speak English may soon be prioritised by police, a top cop has said. (*Sun*, 2017)

In the 2017 data we see a strong association between *speak English* and *patients* (Table 6), and this collocate appears in reports of migrant patients receiving longer medical appointments on account of their (purportedly lacking) English proficiency (38). Similarly, the reporting of Metropolitan Police Deputy Commissioner Craig Mackey's suggestion that under-pressure forces may have to prioritize face-to-face visits for those most vulnerable, which may include people for whom English is not a first language, is reflected in the 2017 collocates *crime*, *victim* and *priority* (Table 6). Both of these stories represent migrants who do not have English as a first language as eroding native-English speakers' access to public services, notably on the basis of their language skills. In (38), such migrants are implicated in discussions of struggling GP practices, "soaring patient numbers" and the fact that people "cannot get appointments", with no mention of other alternative influences such as an ageing UK population, children turning 16 and receiving their National Insurance numbers or, perhaps most crucially, a real-term decrease in Government funding of healthcare services (cf. Taylor 2013). Similarly, in the context of the policing story, there is no acknowledgement that the very need for prioritization could itself be a result of the reduction in police numbers and Government funding.⁵ Regardless of how true it is that people who do not have a native competency in English are more

5. According to a recent report by the charity Full Fact, taking inflation into account, overall funding for the police fell by 19% between 2011 and 2019. By comparison, police funding increased by 31% between 2000 and 2011, see <<https://fullfact.org/crime/police-funding-england-and-wales/>>.

vulnerable than those who do, and require longer appointment times, such groups may always be considered as undeserving recipients of any re-allocation of public services in an enduring media narrative in which they are framed as the Other.

The crosshairs of right-leaning media attention regarding the position of non-native English speakers in the public and private sectors have clearly shifted over the last decade. Consistent reporting on the perceived danger of migrant doctors and nurses to patients' health, including calls for their sacking, subsided after legislative changes which demand "fluent" English for such roles. Since then, the focus has shifted to non-native English-speaking workers in the private sector, namely care homes, waiters and taxi drivers, possibly with a view to precluding (some) non-native English-speaking migrants from working in these occupations in the same way as they were, by this point, prevented from taking up certain jobs in the public sector without English language qualifications. At the same time, non-native English speakers' access to public services has come under scrutiny, highlighting areas where they are perceived to have received privileged access to those services. These may be the first stages of the right-leaning media initiating very real public and policy discussions about the distribution of the most important public services on the basis of native language and/or language proficiency, wherein the indigenous British in-group are competing with the migrant out-group. Such divisive, exclusionary and prejudiced media recommendations would appear unambiguously racist were it not for their being disguised as discourse about language, in a mediated process which, Lippi-Green (2012: 74) argues, "props open a back door" to discrimination. In Skutnabb-Kangas's (1988: 13) terms, this legitimization of unequal division of resources between groups on the basis of their language amounts to "linguicism", an "analogous concept to racism, sexism, classism etc."

4. Conclusion

The study reported in this chapter has found that although non-native English-speaking migrants have been consistently represented in negative and stigmatizing ways by the right-leaning British press, this representation has not been homogenous in nature over the 13 years between (and including) 2005 and 2017. On one hand, we have seen the gradual broadening of social actors reported as not "speaking English", from exceptional cases such as professional sportspeople, through to the general and aggregated "immigrants". Similarly, we have seen English shift from being one language in migrants' multilingual repertoire, to the entrenchment of a deficit model of linguistic competence in which their other languages are backgrounded and dismissed as unimportant. Over the same period, we have seen the responsibility of integration shift from Government to migrants, to

the extent that migrants' learning English is presented as the sole remedy for fixing communities that have reportedly already broken down. Finally, we observed a shift in focus from non-native English speakers *working* in public service roles to those *using* public services, while ideologies relating to linguistic competence and suitability for work are being smuggled into the private sector via the well-trodden route of healthcare and the reporting of language-related events from other countries. Throughout our analysis, we have, where possible, linked these changes in representation to changes in UK society, whether that be legislative changes, shifts in public opinion, or the 2011 Census. In Wright and Brookes (2019), we viewed the Census and its results as a watershed moment in the speak English debate. However, taking a broader view over time, it has become clearer that this may not necessarily be the case. Rather, much like the right-leaning media reporting we have examined, the Census itself may be a symptom of the social and political climate of the time (Sebba 2017). In fact, the scale and nature of press reporting may have contributed to the very inclusion of the language-related questions in the Census. While the media clearly has a dialectal relationship with society – with each shaping the other (Fairclough 1995) – it is not easy to determine the extent to which media discourses around the speak English debate have indeed instigated and influenced the various social, political and legislative events mentioned throughout this chapter, or indeed the other way around.

We have observed in this analysis that the speak English debate, as we have called it, is a dynamic one and one in which the scope and targets are broadening. Whether it is the shift from elite sportspeople to migrants generally, those in the public sector to those in the private sector, and from workers to customers, the number of non-native English speakers being implicated in the debate is increasing. Similarly, changes in representations are such that schoolchildren were initially a burden to schools and English-speaking children, before becoming a threat to the native ingroup, and migrants have gone from being encouraged and supported in their integration to local communities to being blamed for the breakdown of such communities. Over time the debate has permeated new areas of life and new groups in Britain, all the while normalizing the negative and stigmatizing ideologies and representations we have described. In doing so, the right-leaning media make it easier to recommend, justify and rationalize exclusionary and discriminatory public opinion and government policy.

To conclude this chapter on a methodological note, the approach we have adopted in this study demonstrates the value of extending previous corpus studies to cast light on discourses that may otherwise have been assumed to be static (Baker & McEnery 2019). In Wright and Brookes (2019) we identified a number of representational patterns surrounding non-native English speakers *after* the 2011 Census. However, by including in our analysis the six years preceding the Census,

and utilizing collocation as a means to identify and monitor pertinent representations over time, we have found that while some representations are fairly consistent, others are intimately time-bound. At the same time, though, it is clear that there are a wider range of socio-political conditions than the Census that have influenced the press representations of non-native English speakers. Thus, our approach also serves to reiterate the importance of fully integrating corpus techniques within a CDA framework. It is only through viewing and resituating our texts within their original temporal, social and political contexts that we have been able to anchor our analysis to relevant moments in time, enriching our interpretation of the representational patterns revealed by the corpus data.

References

Corpora and tools

Nexis = <<http://www.lexisnexis.com/uk>>

WordSmith = Scott, Mike. 2016. *WordSmith tools version 7*. Stroud: Lexical Analysis Software.

Other references

- Baker, Paul, Gabrielatos, Costas, KhosraviNik, Majid, Krzyżanowski, Michał, McEnery, Tony & Wodak, Ruth. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3): 273–306. <https://doi.org/10.1177/0957926508088962>
- Baker, Paul, Gabrielatos, Costas & McEnery, Tony. 2013. *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge: CUP.
- Baker, Paul & McEnery, Tony. 2015. Who benefits when discourse gets democratised? Analysing a Twitter corpus around the British Benefits Street debate. In *Corpora and Discourse Studies: Integrating Discourse and Corpora*, Paul Baker & Tony McEnery (eds), 244–265. London: Palgrave.
- Baker, Paul & McEnery, Tony. 2019. The value of revisiting and extending previous studies: The Case of Islam in the UK Press. In *Quantifying Approaches to Spoken Discourse for Social Scientists*, Ronny Scholz (ed.), 215–249. London: Palgrave Macmillan.
- Balmas, Meital. 2017. Bad news: The changing coverage of national leaders in foreign media of western democracies. *Mass Communication and Society* 20(5): 663–685.
- Bednarek, Monika. 2016. Voices and values in the news: News media talk, news values and attribution. *Discourse, Context and Media* 11: 27–37. <https://doi.org/10.1016/j.dcm.2015.11.004>
- Bell, Allan. 1991. *The Language of News Media*. Blackwell: Oxford.
- Billig, Michael. 1995. *Banal Nationalism*. London: Sage.
- Blackledge, Adrian. 2002. The discursive construction of national identity in multilingual Britain. *Journal of Language, Identity and Education* 1(1): 67–87. https://doi.org/10.1207/S15327701JLIE0101_5

- Blackledge, Adrian & Creese, Angela. 2009. 'Because tumi Bangali': Inventing and disinventing the national in multilingual communities in the UK. *Ethnicities* 9(4): 451–476. <https://doi.org/10.1177/1468796809345605>
- Bourdieu, Pierre. 1998. *On Television and Journalism*. London: Polity Press.
- Brezina, Vaclav, McEnery, Tony & Wattam, Stephen. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2): 139–173. <https://doi.org/10.1075/ijcl.20.2.01bre>
- Cook, Guy. 2010. *Translation in Language Teaching*. Oxford: OUP.
- Fairclough, Norman. 1989. *Language and Power*. London: Arnold.
- Fairclough, Norman. 1995. *Media Discourse*. London: Arnold.
- Forrest, Adam. 2018. Woman 'brutally punched' for speaking Spanish in racist attack on London Overground train. *The Independent*, 27 October 2018. <<https://www.independent.co.uk/news/uk/crime/woman-punched-racist-attack-london-overground-speaking-spanish-tfl-police-a8604436.html>> (7 November 2019).
- Gablasova, Dana, Brezina, Vaclav & McEnery, Tony. 2017. Collocations in corpus-based language learning research: Identifying, comparing and interpreting the evidence. *Language Learning* 67 (Suppl. 1): 155–179.
- Gabrielatos, Costas & Baker, Paul. 2008. Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005. *Journal of English Linguistics* 36(1): 5–38. <https://doi.org/10.1177/0075424207311247>
- Hill, Amanda. 2017. 'Hostile environment': The hardline Home Office policy tearing families apart. *Guardian*, 28 November 2017. <<https://www.theguardian.com/uk-news/2017/nov/28/hostile-environment-the-hardline-home-office-policy-tearing-families-apart>> (7 November 2019).
- Kelly-Holmes, Helen & Milani, Tommaso M. 2011. Thematising multilingualism in the media. *Journal of Language and Politics* 10(4): 467–489. <https://doi.org/10.1075/jlp.10.4.01kel>
- Lippi-Green, Rosina. 2012. *English with an Accent: Language, Ideology and Discrimination in the United States*. London: Routledge.
- McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: CUP.
- Piller, Ingrid. 2001. Naturalization language testing and its basis in ideologies of national identity and citizenship. *International Journal of Bilingualism* 5(3): 259–277. <https://doi.org/10.1177/13670069010050030201>
- Potter, Jonathan, Wetherell, Margaret & Chitty, Andrew. 1991. Quantification rhetoric – cancer on television. *Discourse & Society* 2(3): 333–365. <https://doi.org/10.1177/0957926591002003005>
- Sebba, Mark. 2017. 'English a foreign tongue': The 2011 census in England and the misunderstanding of multilingualism. *Journal of Language and Politics* 16(2): 264–284. <https://doi.org/10.1075/jlp.14026.seb>
- Skutnabb-Kangas, Tove. 1988. Multilingualism and the education of minority children. In *Minority Education: From Shame to Struggle*, Tove Skutnabb-Kangas & Jim Cummins (eds), 9–44. Bristol: Multilingual Matters.
- Stickings, Tim. 2018. Woman is viciously attacked on late night Tube train for speaking Spanish by two fellow passengers who told her she should 'talking English in England'. *MailOnline* 14 April 2018. <<https://www.dailymail.co.uk/news/article-5615345/Woman-viciously-attacked-late-night-Tube-train-speaking-Spanish.html>> (7 November 2019).
- Taylor, Roger. 2013. *God Bless the NHS: The Truth Behind the Current Crisis*. London: Faber & Faber.

- van Dijk, Teun. 2001. Multidisciplinary CDA: A plea for diversity. In *Methods of Critical Discourse Analysis*, Ruth Wodak & Michael Meyer (eds), 95–120. London: Sage.
- van Dijk, Teun. 2015. Critical discourse analysis. In *The Handbook of Discourse Analysis*, 2nd edn, Deborah Tannen, Heidi Hamilton & Deborah Schiffrin (eds), 466–485. London: Wiley.
- van Leeuwen, Theo. 2007. Legitimation in discourse and communication. *Discourse & Communication* 1(1): 91–112. <https://doi.org/10.1177/1750481307071986>
- Vessey, Rachelle. 2016. Language ideologies in social media: The case of Pastagate. *Journal of Language and Politics* 15(1): 1–24. <https://doi.org/10.1075/jlp.15.1.01ves>
- Vessey, Rachelle. 2017. Corpus approaches to language ideology. *Applied Linguistics* 38(3): 277–296. <https://doi.org/10.1093/applin/amv023>
- Wright, David & Brookes, Gavin. 2019. ‘This is England, speak English!’: A corpus-assisted critical study of language ideologies in the right-leaning British press. *Critical Discourse Studies* 16(1): 56–83. <https://doi.org/10.1080/17405904.2018.1511439>

PART II

Changing language

That's absolutely fine

An investigation of *absolutely* in the spoken BNC2014

Karin Aijmer

University of Gothenburg

This chapter investigates the meanings and functions of the intensifier *absolutely*, both synchronically and diachronically. The research questions are asked against the background of *absolutely* becoming more frequent over a short period of time as shown by a comparison of the BNC1994 and the BNC2014. The results show that the developments undergone by *absolutely* can be described as a trajectory from degree modifier to emphasizer and to discourse marker, as *absolutely* increases in subjectivity. The description of the semantic and pragmatic developments also requires a sociolinguistic analysis of factors such as age and gender. The analysis shows that young speakers use it less often than older speakers and that it was first used by male speakers and only later by females.

1. Introduction

Intensifiers, such as *absolutely*, *really*, and *so*, have an important function in language to convey the speaker's attitudes or emotions. Speakers do not only say that something is "exciting" but it is "absolutely exciting" or "really exciting". Intensifiers have attracted a great deal of scholarly attention because they are in a state of continuous and rapid change; new intensifiers emerge or reappear after a time of decline, they fall out of fashion and are replaced by new, more colourful ones (see e.g. Altenberg 1991; Bolinger 1972; Ito & Tagliamonte 2003; Tagliamonte 2016). The changes undergone by intensifiers can be observed also in a short time perspective (Aijmer 2018; Fuchs 2017; Macaulay 2006; Paradis 2000; Stenström, Andersen & Hasund 2002). The trigger for this study was the observation that *absolutely* seems to be rising in frequency and the possibility to study on-going changes on the basis of present-day spoken corpora from different periods.

Absolutely is an amplifier with the function of scaling properties upwards from an assumed norm to an end-point on a degree scale (Quirk et al. 1985: 590).

Amplifiers can be further subdivided into maximizers (e.g. *completely, totally, utterly, absolutely*) and boosters (e.g. *very, really*). The function of maximizers, exemplified by *amazing* and *brilliant*, is to reinforce extreme adjectives, expressing a superlative degree of a certain property (Paradis 1997: 56). Another type of adjectives modified by maximizers are limit or 'either-or' adjectives such as *certain* or *true*. Boosters mainly combine with scalar adjectives such as *nice, kind, expensive* fulfilling the criteria for gradability (Paradis 1997: 51).

Absolutely can intensify adjectives, adverbs, pronouns, prepositions, determiners, modal auxiliaries, verbs and nouns. Examples (1) to (2) illustrate *absolutely* with a scope over an adjective or a verb. Example (3) shows that *absolutely* can be used in syntactic contexts where it lacks a head word and is no longer an intensifier.

- (1) This counsellor she was *absolutely brilliant* (S28F)¹
- (2) I love my room I *absolutely love* it (SK8T)
- (3) B: you can definitely admire people of the same sex
 A: *yeah absolutely mm yeah*
 B: and I guess be besotted with them without it being sexual (S38V)

However, both synchronically and diachronically, we also need to consider intensifiers from the point of view of the pragmatic work they are carrying out in conversation. Intensifiers can come to function as a vehicle of emphasis and subjectivity, heightening the force of the proposition and marking the intensity of speaker involvement and commitment (Cacchiani 2005; Claridge 2011). *Absolutely* has, for example, also been referred to as an emphasisier reinforcing the truth value of the clause.

The aim of this chapter is to study the different meanings and functions that *absolutely* has and how they are related synchronically and diachronically. The history of *absolutely* is complex and messy reflected in a great deal of fuzziness between its intensifying and emphasizing function. The research questions will be asked against the background of the changes in frequency *absolutely* is undergoing over a short time period. What is the importance of the context for syntactic and semantic change? What are the grammatical and pragmatic mechanisms responsible for the changes? What happens in a short time perspective, and what happens in a longer perspective? How are the distribution and diffusion of *absolutely* related to the age and gender of the speakers?

The chapter is organized in the following way. Section 2 presents the methodological approach and the corpora used. Section 3 introduces quantitative corpus data showing that *absolutely* is becoming more frequent in the new spoken

1. References are to BNC2014 unless stated otherwise.

component of the *British National Corpus* (BNC2014S) than in the traditional BNC1994. Sections 4 and 5 contain the qualitative analysis covering the meanings and functions that *absolutely* can have depending on the syntactic and semantic environment where it occurs. In the discussion in Section 6, I will examine how the changes undergone by *absolutely* can be explained in a grammaticalization perspective. Section 7 contains the conclusion.

2. Methods

There has been a great deal of research on intensifiers in variational sociolinguistics (see e.g. Ito & Tagliamonte 2003; Tagliamonte 2008; Tagliamonte & Roberts 2005). In order to describe variation and change in the system of intensification, the standard methodology has been to study the developments of intensifiers across generations in apparent time. This approach makes it possible to trace change in the usage of intensifiers as reflected in their frequencies and distribution across speakers belonging to different age groups. In addition to such studies we can now also observe change in real time on the basis of diachronic corpora.

The Spoken BNC2014 (Love et al. 2017) is a large corpus of British English created as a resource to study phenomena characteristic of contemporary spoken language. The corpus is designed along the same lines as the older BNC1994, thus making it possible to observe linguistic change taking place over a twenty-year period. Both corpora provide rich metalinguistic data about the speakers' backgrounds and the conversational context. The factors which are important for a variational perspective on intensifiers are above all the age and gender of the speakers.

The size of the Spoken BNC2014 is circa ten million words. In order to analyse *absolutely*, I have used the sample version of five million words which makes the dataset comparable in size with the demographic part of BNC1994. Occurrences of *absolutely* were extracted from both BNC1994DS and BNC2014S using the CQP (Corpus Query Processor) facility (Hardie 2018: 27). It was also possible to use this tool to search for occurrences of *absolutely* restricted to a particular age group or to gender. The instances were then classified with respect to the syntactic and semantic context.

3. The frequency of *absolutely*

Absolutely is more common in spoken than in written English. This has, for example, been shown by Carretero (2010) who used BNC to compare *absolutely* in spoken and written English with its Spanish counterpart *absolutamente*. According

to Carretero (2010: 203), *absolutely* was almost four times as frequent in the spoken part of BNC1994 as in the written component of the corpus (188.90 vs. 42.03 occurrences per million words). The higher frequency of *absolutely* in speech suggests that the changes *absolutely* is undergoing take place in the on-going interaction among speakers.

The present study takes as its starting point the observation that the intensifier *absolutely* is more frequent in the 2010s (BNC2014S) than in the 1990s (BNC1994DS; see Table 1).

Table 1. The frequency of *absolutely* in BNC1994DS and BNC2014*

	Words	N	pmw	LL	Sig level
BNC1994D	4,788,882	671	133.8		
BNC2014S	5,014,655	1,337	279.2	195.52	<0.0001

* On log likelihood, see <ucrel.lancs.ac.uk/llwizard.html>.

Absolutely shows a fairly steep rise from 133.8 tokens per million words in BNC1994DS to 279.2 tokens per million words in BNC2014S.² As can be seen from Table 1, the difference in the frequency of *absolutely* in the corpora is statistically significant. Changes in frequency can also be observed in the distribution of *absolutely* across age groups (see Figure 1). The speakers have been grouped into different age bands on the basis of the classification system introduced in BNC1994.³

Figure 1 shows that there are clear differences in the use of *absolutely* depending on age group. *Absolutely* was more frequent in the speech of adults and older speakers than among adolescents (the age bands 0–14, 15–24). There is a small drop in frequency in the 25–34 age group in BNC1994DS and, in BNC2014S, in the 35–44 age band, but then the rate rises again. The largest difference between the corpora is the rise in frequency of *absolutely* with the oldest speakers in BNC2014S. The differences are statistically significant in all the groups except the 15–24-year olds (see Appendix 1).

Other studies have also reported differences relating to age although the results are difficult to compare with the data from BNC. Núñez-Pertejo and Palacios Martínez (2014: 221) found that *absolutely* was four times more frequent in the language of adults than in teenage language. The data consisting of adolescent speech was extracted from the COLT Corpus (*The Bergen Corpus of London Teenage*

2. The data are based on BNC1994 XML-edition and not on BNC1994 World Edition which was used by Carretero (2010). The frequencies per million words are therefore slightly different.

3. When the exact age of the speakers in BNC2014 was not known the example has been classified as (age) unknown.

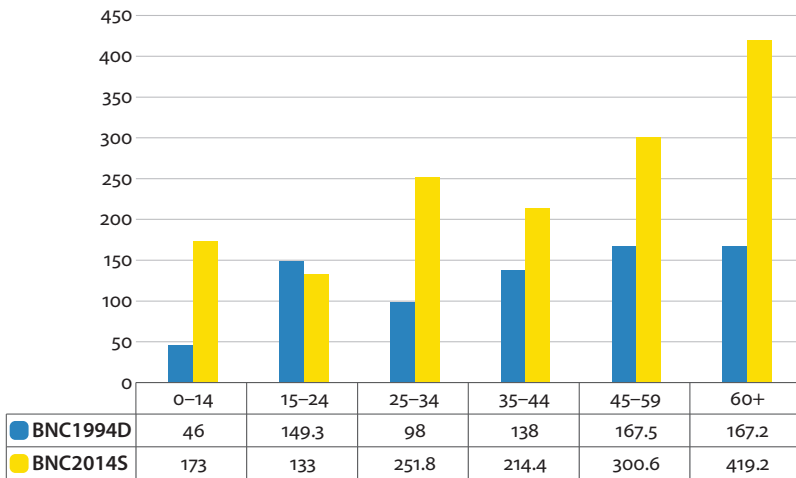


Figure 1. The frequency of *absolutely* with speakers belonging to different age groups in BNC1994DS and BNC2014S (pmw)

Language) supplemented with occurrences from the *Saarbrücken Corpus of Spoken English*. A comparison was made with adult data from *The Diachronic Corpus of Present-Day Spoken English* (sampled from the *London-Lund Corpus* and the British component of the *International Corpus of English*). The results from Núñez-Pertejo and Palacios Martínez (2014) agree with the findings in BNC2014S in that older speakers were the most frequent users of *absolutely*. Murphy (2010) was also interested in studying the influence of age and gender on language and collected a small corpus (*Corpus of Age and Gender*) of approximately 90,000 words with this in mind. The subjects were selected according to their chronological age matching ‘young adulthood’ (20–29-year-olds), ‘mid-adulthood’ (40–49-year-olds) and ‘old adulthood’ (70–90-year-olds). *Absolutely* was shown to be used differently depending on the age and gender of the speakers. The analysis pointed to a cut-off point between young speakers in their twenties and the 40-year-olds (the most frequent users). This again agrees with my data which show that speakers started using *absolutely* more frequently when they reached middle age (the 45–59 age-band in BNC2014S). However, in Murphy’s study, older speakers used *absolutely* less frequently (see Table 2, based on Murphy 2010: 147–148). The reason may be that the older speakers in her data were all in their seventies or eighties where the frequency of *absolutely* is also lower in BNC2014S.⁴

4. This is apparent from the more detailed age classification in BNC2014.

Table 2. The distribution of *absolutely* across speakers in different age groups according to Murphy (2010) (pmw)

20s	40s	70s/80s
414	803	108

Gender is another factor which has been shown to have an impact on the frequency and use of intensifiers. Figure 2 shows that, in BNC1994DS, *absolutely* was used more frequently by male speakers in all the age groups except among the 60+ speakers who use it less often. The differences are significant in all the age bands with the exception of the 35–44-year-olds (see Appendix 2).

In BNC2014S (see Figure 3), female speakers using *absolutely* were more frequent in all age groups with the smallest difference among the 25–34-year-olds. Female speakers used *absolutely* frequently particularly in the older age groups.

If we cross-tabulate age and gender as in Figure 3, it is shown that female speakers in the older age groups use *absolutely* more than male speakers in their speech. The difference is statistically significant at the $p < 0.0001$ level (see Appendix 3). The higher frequency of *absolutely* in the speech of older female speakers agrees with Murphy’s (2010) findings that *absolutely* was never used by male speakers in their seventies and eighties to be compared with 108 times by the female speakers in this age group.

Gender needs to be highlighted as an important sociolinguistic factor associated with the frequency and change of intensifiers. However, there is no consensus

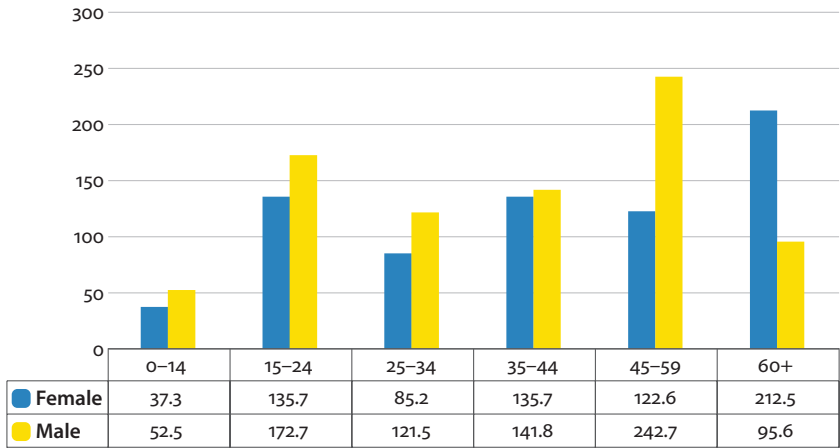


Figure 2. The frequency of *absolutely* used by female and male speakers in different age groups (pmw) in BNC1994DS

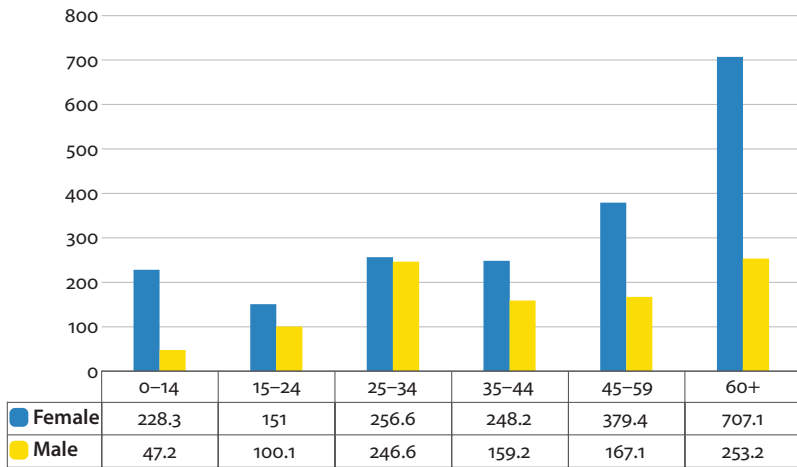


Figure 3. The frequency of use of *absolutely* with female and male speakers in BNC2014S (pmw)

about the role played by gender. It has been claimed by scholars both that men use intensifiers more than women and that intensifiers are preferred by women (Hessner & Gawlitzek 2017). Fahy (2002) has, for instance, shown that male speakers used amplifying intensifiers to a higher degree than female speakers in a corpus consisting of transcripts from a computer conference. In a more recent study, Xiao and Tao (2007) found that male speakers used more amplifiers than female speakers in the spoken component of BNC1994 although there was only a slight difference between the speakers. The reverse has also been found: using data drawn from BNC1994 and BNC2014S, Fuchs (2017) found evidence that men overall used fewer intensifiers than women and that this situation had not changed over time.

One explanation for the different findings may be that that men and women have different preferences for individual intensifiers (Hessner & Gawlitzek 2017: 423). Hessner and Gawlitzek's investigation of female and male speakers' usage of intensifiers on the basis of BNC2014S showed that out of 39 intensifiers observed, the female speakers used 18 intensifiers more often than men and that the higher frequency in female speech was significant. Stenström et al. (2002) found that intensifiers on the whole were more frequent in girls' than in boys' conversations in the COLT corpus. However, Stenström et al. (2002) also found that the boys tended to use the strongest ones – such as *extremely*, *completely*, *absolutely* – more than girls, suggesting that there may be differences depending on the intensifier.

4. The frequency of *absolutely* with different syntactic head words

Many studies of intensifiers (e.g. Ito & Tagliamonte 2003; Núñez-Pertejo & Palacios Martínez 2014; Palacios Martínez & Núñez-Pertejo 2012; Stenström et al. 2002; Xiao & Tao 2007) only investigate occurrences where intensifiers co-occur with adjectives. However, *absolutely* collocates with a number of different heads. In Table 3, the occurrences have been categorized with regard to the syntactic element modified by *absolutely* (a word or a clause). The comparison in Table 3 is based on a random sample of 500 instances of *absolutely* from each corpus.

Table 3. The frequency of collocations of *absolutely* with different syntactic head words

	BNC1994DS		BNC2014S	
	N	%	N	%
Occurrences with adjectives	229	45.8	146	29.2
Occurrences with verbs	33	6.6	61	12.2
Occurrences with past participles*	50	10.0	37	7.4
Occurrences with <i>no, nothing, all, everything, anything</i>	25	5.0	33	6.6
Occurrences with preposition phrases	4	0.8	7	1.4
Occurrences with nouns	14	2.8	9	1.8
Clause-final discourse marker	20	4.0	23	4.6
Response marker	92	18.4	147	29.4
Other**	33	6.6	32	6.4
Total	500	100.0	500	100.0

* The examples with past participles will be discussed together with adjectives in Section 4.1.

** Instances where *absolutely* was difficult to classify, for example because it was interrupted.

The following discussion will be based on the functions that *absolutely* has with different head words. *Absolutely* will first be analysed when it collocates with adjectives.

4.1 *Absolutely* with adjectives

In BNC1994DS, *absolutely* modified an adjective in almost half the instances (45.8%),⁵ compared with 29.2% in BNC2014S. The short-time change undergone by *absolutely* with adjectives over 20 years can be described in terms of a decline in frequency both with regard to the type of adjective and the number of tokens (see Figure 4).

5. In Carretero (2010: 204), the figure is 55.6%.

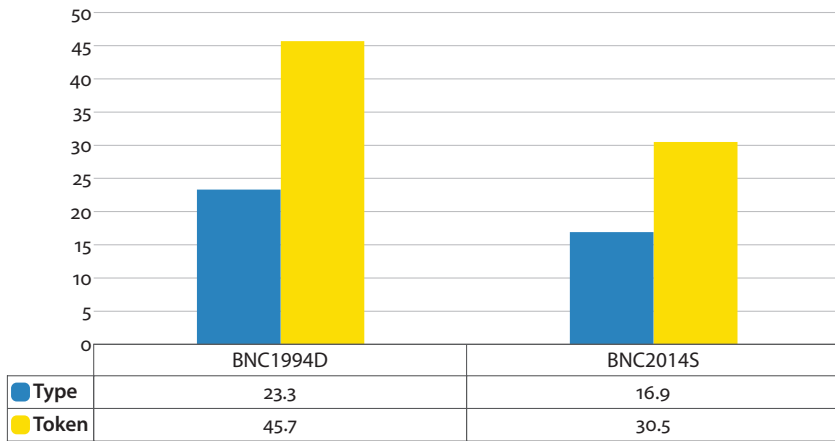


Figure 4. The frequency of *absolutely* plus adjective in BNC1994DS and BNC2014S (types and tokens; pmw)

Absolutely has several intensifying uses depending on the collocating adjective. In the default case, *absolutely* is a maximizing intensifier, reinforcing the utmost degree of an adjective on a scale. The most frequent adjectives (found in both corpora) are polarized or extreme adjectives expressing an emotion. Examples are *amazing*, *appalling*, *beautiful*, *disgusting*. The collocating adjective can be negative, as in *absolutely disgusting* and *absolutely horrible*, or positive, as in *absolutely brilliant* or *absolutely marvellous*. There were more negative adjectives in BNC1994DS (117 negative; 90 positive), while in BNC2014S there was hardly any difference (77 negative; 74 positive).

The evaluative adjectives can be colourful and dramatic as in the combinations *absolutely retarded* or *absolutely mental* which are age-related and can be used to indicate group identity:

- (4) A: oh man if that had been me I would have gone absolutely ape shit
 B: yeah
 A: *absolutely mental* (S6J2)

Absolutely mental is preferred to the common and old-fashioned ‘very stupid’ because the combination “warrants greater attention” (Swales & Burke 2003: 13). Instances of informal and dramatic adjectives or participles (from the BNC2014S) used by adolescent speakers (mainly 15–24-year-olds) are *devastated*, *farcical*, *humungous*, *mental*, *retarded*, *schiz*, *wasted*, *whacked*, *vile*. There were also differences between age groups. The 60-year-olds used *absolutely blottoed* to be compared with *absolutely wasted* chosen by younger speakers to describe the same situation.

Absolutely is sensitive to contextual factors and can be a booster as well as a maximizer. When *absolutely* modifies scalar or gradable adjectives, such as *beautiful* or *tired*, it functions as a booster. It means something like *really* which also allows a shift towards intensification with degree adjectives (Paradis 2003). *Absolutely* can also be used to assert the speaker's claims or opinions with adjectives related to certainty (BNC1994DS *certain, clear, convinced, correct, sure, true*; BNC2014 *necessary, right, true, wrong*).

Absolutely combines with other intensifiers to express subjectivity and speaker involvement. These occurrences reflect the speaker's strong feelings in the speech situation (annoyance, disgust, etc.) but are not associated with a particular age group. In Example (5), the speaker is a female language teacher in her thirties:

- (5) and she said she was *absolutely* (.) *you know really* traumatised by it (S33B)

Absolutely was also used by young speakers in more creative combinations with intensifiers. In (6), the speaker is a 27-year-old female:

- (6) every time we visit them we go out the night before and get *absolutely fucking brutally obliterated* and then the next day have like a really awesome hangover (SDJA)

The extent of expressivity can also be indicated by the intensifier's collocations with adjectives associated with hyperbole. In Example (7), the speaker has used *absolutely* with "freezing cold" rather than with "cold" alone for exaggeration and emphatic effect:

- (7) mm hm (.) the only time when you would find having a car useful is in um when it's chucking it down with rain or it's *absolutely freezing cold* (SJDM)

In (8), "crystal clear" has been used because it is more expressive than the simple "clear":

- (8) I thought it might be erm a bit er dirty or smelly but it was *absolutely crystal clear* and you could see all the plants sparkling through the water (SAXQ)

Absolutely with adjectives varies in its distribution and use across age groups as shown in Figure 5. The results show an increase of *absolutely* in the speech of the oldest age group (statistically significant at the $p < 0.0001$ level; see Appendix 4). The difference between the corpora is also significant when we compare *absolutely* in the speech of the 15–24-year-olds and the 35–44-year olds (statistically significant at the $p < 0.005$ level).

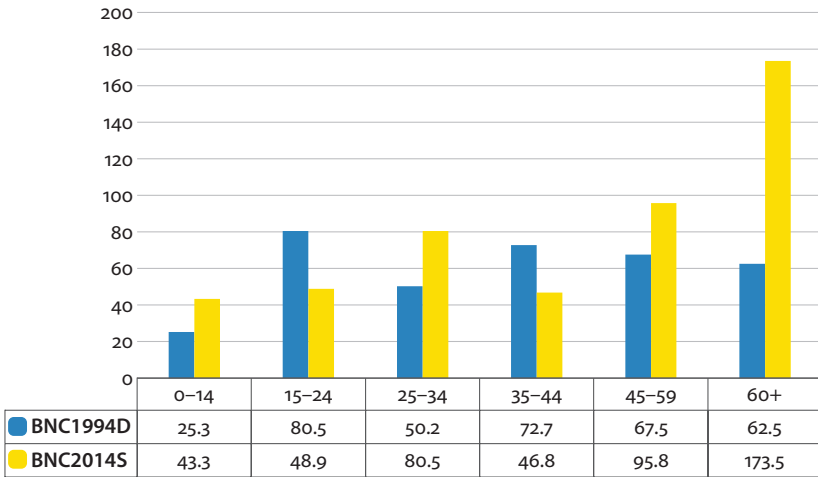


Figure 5. The frequency of *absolutely* plus adjective in the spoken language of different age groups in BNC1994DS and BNC2014S (pmw)

The high frequency and rise in the number of adjectives modified by *absolutely* with older speakers is confirmed by the observations of other scholars. Núñez-Pertejo and Palacios Martínez (2014: 221) found, for example, that adults generally used adjectival modifiers with *absolutely* more often than adolescents (adults 56.5%; adolescents 44.3% of the total number of instances).

4.2 *Absolutely* modifying a verb

The proportion of *absolutely* with verbs is higher in BNC2014S than in BNC1994DS (12.2% vs. 6.6%). This makes it likely that a development has taken place sequentially from adjectives to verbs as head words. The question is therefore in what contexts the semantic development takes place and how new functions arise. The occurrences with verbs illustrate that intensification does not always involve degree, but *absolutely* can be used with other elements than adjectives as an emphaser. The emphaser function has been defined as follows: “Emphasizers reinforce the truth value of the clause they apply to and can be related to modality distinctions expressed at the clause level” (Nevalainen & Rissanen 2002: 361).

A first observation we can make is that *absolutely* occurs with accomplishments (telic situations) and achievements with emphaser function. In (9), the verb intensified by *absolutely* is *leap at* referring to a verbal activity for which telicity is relevant (cf. also *absolutely hit me*, *absolutely knocked her sideways*, *absolutely did*

my head in taken from BNC2014S). In these expressions, *absolutely* has the effect of highlighting an event (often of a negative kind) where an endpoint is reached or overreached.

- (9) and one of the drunken American guys decided it would be a good idea to pull a moonie at them and so he just work- went stand in front of them pulled his trousers down in front of them and *they just absolutely leapt at him* (SJT)

Absolutely is also found with verbs conveying some suggestion of exaggeration or emotional expressiveness in their predication which can be further emphasized by the intensifier (Quirk et al. 1985: 685). The most common verbs with *absolutely* were stative verbs meaning *love* and *hate*. *Love* and *hate* can be regarded as more expressive or even hyperbolic in comparison with the weaker *like* and *dislike*. The loving and hating verbs make up about 36% of the verbs in BNC2014S, and 21% of the instances in BNC1994DS (see Examples (10)–(12)). Related but less frequent verbs in the same semantic field are illustrated by *adore* (both BNC1994DS and BNC2014S), *thrive on*, *detest*, *loathe*, *cannot stand*.

- (10) it's such a novelty for them they *absolutely love* it (S584)
 (11) it was horrible I hate mice (.) I *absolutely hate* mice (SC7H)
 (12) I *absolutely hate* her (.) like totally (S7V3)

Other types of verbs emphasized by *absolutely* can also be characterized by subjectivity and exaggeration. In Example (13), the speaker has chosen to say “reek with garlic” rather than “smell of garlic” to strengthen the negative connotations associated with the smell.

- (13) You see. Jim doesn't like. I mean. we'd go to Carriages probably cos it's a a proper Italian Yeah. that *absolutely reeks with garlic* and all the rest of it.
 (BNC1994DS, KBD)

Obsessed with and *ridden with* are similar instances associated with negative connotation and an emphasizing function.

Absolutely is also found with progressives having a subjective function in addition to their aspectual use. According to Wright (1994: 480), subjective progressives serve to “express momentary but intense experiences; experiences which are, moreover, cognitive or emotional, rather than physical” (quoted from Killie 2004: 27). *Absolutely* has epistemic meaning (‘definitely’) increasing the strength and intensity of the described experience. In (14), the progressive form “gagging for a cigarette”, for example, indicates subjectivity and not aspect. *Absolutely* plays an important role since it lends itself to hyperbole:

- (14) I don't know (.) I'm starving (.) and I'm *absolutely gagging for a cigarette* (S33N)

The verb phrase often has an idiomatic character:

- (15) her friends were just lit- like *absolutely like taking the Mickey out of her* (SPXX)

In (16), the reference is to a physical rather than emotional experience:

- (16) it was *absolutely pissing down* (S3RZ)

BNC1994DS has “absolutely bumbling with life” as the only example in this category.

The verbal contexts in which *absolutely* occurs also include auxiliaries (typically *must, have got to, have, be, do, can*). In (17), the speaker has used *absolutely* because it reinforces the deontic modal meaning of obligation thus making it possible for the speaker to commit herself more strongly to the verbal action:

- (17) sorry I (.) there’s somebody else I *absolutely must* talk to (.) (S6JP)

Absolutely conveys that the speaker has something important (and positive) that she must talk to somebody else about.

With verbs it is sometimes difficult to tell if *absolutely* emphasizes the verb alone or whether the focus is on the assertion and illocutionary force. In (17), *absolutely* could emphasize only the auxiliary (*must talk to*) as in the interpretation suggested, but *absolutely* could also be an emphasizer having the whole clause in its scope (‘it is absolutely the case that I must talk to someone’). *Absolutely* can be paraphrased as an epistemic expression (‘it is certain that I must’).

In (18), *absolutely* modifies the emphatic *do*:

- (18) A: ah will you survive? I’m I’m
 B: >> yeah
 A: tempted to take my socks off now so I can feel the fur
 B: do
 A: do you think I should?
 B: >> *absolutely do* (SGAW)

Speakers’ age can be assumed to be a factor related to the frequency with which *absolutely* is used with verbs. As shown in Figure 6, the increase in frequency over time with verbs is due primarily to older speakers (45–59, 60+). The difference between the speakers is statistically significant at the $p < 0.0001$ level (see Appendix 5). The difference between the speakers is also statistically significant when the speakers using *absolutely* are in the 25–34 age band (significant at the $p < 0.01$ level).

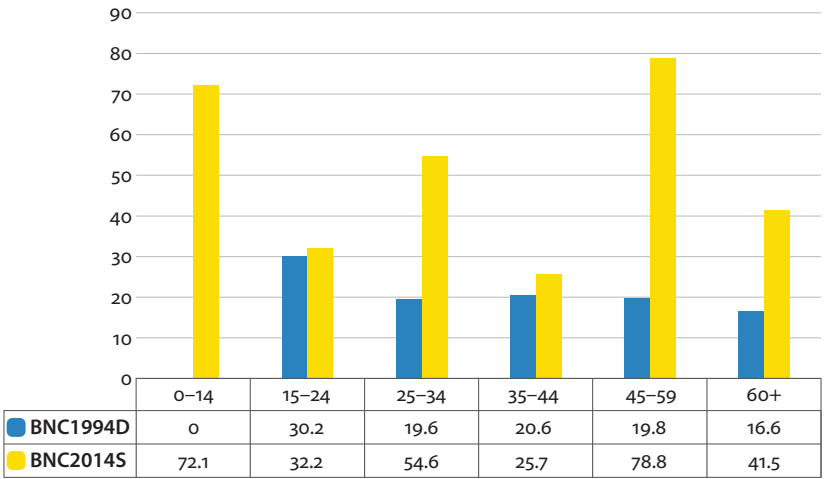


Figure 6. The frequency of *absolutely* before verbs in the speech of different age groups in BNC1994DS and BNC2014S (pmw)

4.3 *Absolutely* followed by *everything*, *all*, *never*, *nothing*, *no*

Absolutely was also used with quantifying expressions such as *everything* or *nothing* to express the speaker’s personal involvement and emphasis. In (19), the speaker does not mean to say that she will learn everything, but “everything” is hyperbolic or subjective and can be further emphasized by *absolutely*:

- (19) and eh I’d recently bought Minecraft so I decided I’d go for that and just learn
absolutely everything about it (SCJL)

Similarly, in (20), it is unlikely that the speaker will eat “anything for breakfast” but *anything* is understood as an exaggeration further emphasized by *absolutely*:

- (20) I could eat anything for breakfast (.) I could eat *absolutely anything* for breakfast (S9CH)

Absolutely was also frequent before *nothing* (13 instances) to express hyperbole, as in (21).

- (21) the Co-op is wonderful they have *absolutely nothing nice* in there at all (S2ZU)

4.4 *Absolutely* with nouns

Crap, *shit*, *ace* represent a case of gradience between noun and adjective (Denison 2001). *Crap* has, for instance, moved from noun to adjective as shown by its co-occurrence with *absolutely* (as in (22)).

- (22) I think it's Claudia Roden's recipe is is *absolutely crap* (SJHL)

Ace has positive connotations:

- (23) Yeah Brilliant. it was all about erm politics. about the struggle for power Oh yeah It was *absolutely ace* it was. it was erm a drama but very very well written (BNC1994DS, KPA)

Other nouns get adjectival properties when they are modified by *absolutely* and could be replaced by adjectives (as in (24)).

- (24) You're right. Imagine in December it would be *absolutely murder*. (BNC1994DS, KBF)

The other instances of nouns are *ages*, *billions*, *tonnes*, *miles away*, *hours*. These nouns do not have a precise meaning which makes them suitable for emphasis (as in (25)).

- (25) I read one and it took me *absolutely ages* to read it I've I've got the others to read but I just sort of (S38F)

The speaker exaggerates the amount of time it took to read the book by using the noun *ages* for hyperbolic effect.

4.5 *Absolutely* before preposition phrases

Preposition phrases can also be emphasized by *absolutely*. In (26), the speaker has chosen to express herself in a figurative way in order to draw attention to the situation where the girls were giggling at themselves doing aquaft.

- (26) ANONnameM and I were giggling away to ourselves (.) er *absolutely on another planet* (.) and they kept looking over and smiling at us like (SVR6)

“(To be) on another planet” is itself hyperbolic and easily co-occurs with *absolutely* for stronger emphatic force.

4.6 *Absolutely* as a discourse marker at the end of the clause

Discourse markers are placed outside the syntactic structure and they do not have propositional meaning (Beeching 2016). Their function is to relate discourse units. However, there is little agreement about the definition of discourse marker and what to include in that category. *Absolutely* at the end of the clause can be regarded as a discourse marker since it is syntactically independent, but its function is emphasizing (as in (27)).

- (27) ANONnameM it is your favourite joke and erm (.) *it's a beautiful day absolutely*
 (.) t- gorgeous spring day no clouds (.) my phone's going (SGJP)

In (28), the topic discussed is a girl who is short but is seeing a boy who is six feet tall.

- (28) A: she can only be five foot four (.) and um I've told her it's act- that should
 be against the law for anyone her height to have anyone who's over six foot (.)
 cos you've gotta leave the tall ones for the tall girls
 B: *for the tall girls (.) absolutely*
 A: *it should be absolutely* (SDR9)

Speaker B uses *absolutely* to assert emphatically that you should leave the tall boys for the tall girls. In A's next turn, *absolutely* emphasizes that 'it should be against the law' for a short girl to have a boy who is over six foot tall.

5. The response marker *absolutely*

Absolutely can also be regarded as a discourse marker when it functions as a response to previous discourse. According to Tao (2007), the use of *absolutely* as a response marker can be attributed to its epistemic meaning ('unquestionably, there is no doubt that') which makes it possible to use it to express strong agreement:

- (29) A: erm just look at the people who've died it's it's miniscule *absolutely*
 miniscule
 B: well that
 A: but look at the money they've made they've brought in injections to every
 single person on this planet now
 B: >> *oh yeah absolutely yep*
 A: at birth or they you know flu vaccines and
 B: >> they try and put yeah they do pregnant women don't they? (S5B4)

When *absolutely* interrupts the preceding speaker, who continues speaking after the intervention, it functions as a backchannel. Backchannels could also be seen as having some of the same characteristics as discourse markers since they are placed outside the clause structure and have a responsive function (McCarthy 2003: 44). Backchannels are associated with active listenership; the hearer uses *absolutely* together with other elements to signal understanding, commitment and involvement (as in (30)).

(30) >> oh it oh it's yeah

A: >> an element of that of course

B: ye-

B: *ye- yeah yeah absolutely absolutely yeah*

A: >> I mean there is usually a correlation between the two er I er I used to do that sometimes correlate it (SJN6)

Absolutely was frequently found as a response to a tag question (14 occurrences in BNC2014S) and it occurs after *(do) you see* and *you know what I mean*, as in (31).

(31) A: *you know what I mean?*

B: *absolutely*

A: and she just wasn't getting it (.) not at all (S7JG)

In (32), *absolutely* is used with the meaning of confirmation after the speakers have come to a (final) agreement about the forms to fill out. *(Oh) absolutely* allows the speaker to show enthusiasm and involvement:

(32) A: oh yes what fun *I'll pass you the appropriate forms to fill out*

B: *oh absolutely yeah yeah*

A: smashing thank you

B: >> cheers thanks – ANONnameM

A: cheeri- bye (SJN6)

In (33), *absolutely* as a response to the message in the preceding clause is in variation with *completely*:

(33) B: then you add up (.) a thousand here two thousand there

A: >> yeah

B: five thousand here twenty thousand for insurance whatever

A: >> *yeah oh absolutely*

B: get up to thirty-six without blinking

A: *completely*

B: and so when erm (.) people like – ANONnameM think do you know what? we'll take it back and we'll run it ourselves (.) they don't reckon on the fact that to stand still they have to raise thirty-six thousand (SNXL)

Almost 30% of the instances of *absolutely* occurred alone with a response marker in the 500-word sample in BNC2014S. In the BNC1994DS, the corresponding percentage was 18.4% indicating that *absolutely* has increased its frequency in relation to its other uses over time. A more detailed study shows that the response marker was used differently by speakers depending on age (see Figure 7).

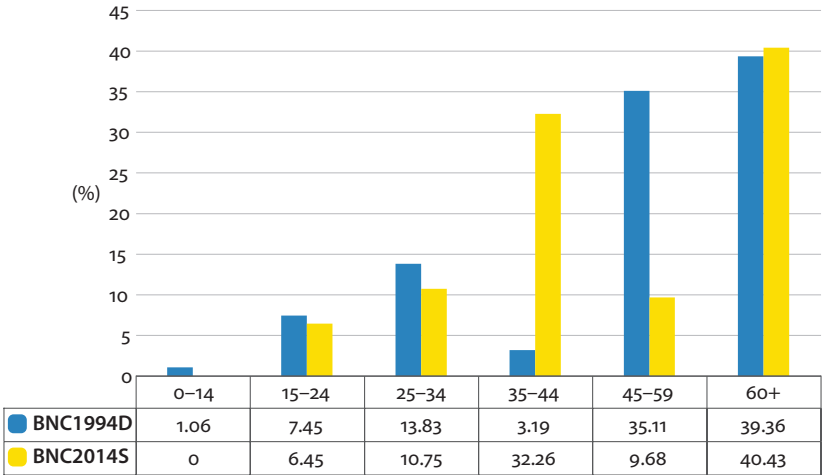


Figure 7. The response marker *absolutely* in the speech of different age groups as a percentage of absolutely occurrences in BNC1994DS and BNC2014S

In BNC1994DS, *absolutely* as a response marker was frequent in the 45–59 age group and among the 60+ speakers. In BNC2014S, it was most frequently used by speakers in the 35–44 age group and by the 60+ speakers.

We can make a comparison with other studies of *absolutely* in this function. In Tao’s (2007: 21) study of *absolutely* (based on American English data) “1/3 of the tokens of *absolutely* or 35% no longer have modifier role” and are functioning as a discourse marker. Núñez-Pertejo and Palacios Martínez (2014: 222) found that the most frequent use of *absolutely* occurred in adult speech, “a function which is hardly allowed at all in the speech of adolescents.” The authors suggest that this could be explained by the fact that adolescents use alternative forms, such as *yeah*, *I know*, *cool*, *OK*. Murphy (2010) studied the use of *absolutely* by female speakers over different age groups and found that *absolutely* was used only once as a response marker by female speakers in their twenties. In the forties age category, women used *absolutely* most often as a response marker, reinforcing the author’s claim that “the 40s boost less and are more concerned with agreeing with each other to keep channels of communication pleasant and to be seen favourably by the other” (Murphy 2010: 147). In the same study, there were no tokens of *absolutely* as a boosted response marker in the oldest group of women (70–90 year old) (ibid.). However, as pointed out above, Murphy did not investigate the use of *absolutely* by speakers in their sixties.

6. Discussion

Degree adverbs have been described “as the most active class of words in terms of undergoing change” (Tao 2007: 223). This raises the question about the mechanisms for change and whether grammaticalization can explain the semantic and pragmatic developments of *absolutely*. Grammaticalization is commonly defined as “the process whereby lexical material in highly constrained pragmatic and morphosyntactic contexts is assigned grammatical function, and once grammatical, is assigned increasingly grammatical, operator-like function” (Traugott 2003: 645).

The grammaticalization of *absolutely* from a manner adverbial to an intensifier has been discussed by Partington (1993). According to Partington (1993: 181), *absolutely* started out as an adverbial with the semantic content “in a manner detached from other things, separately, independently. It also had a modal sense ‘without doubt or condition... certainly, positively’.” The latter meaning is illustrated (34) (quoted from Partington 1993: 181).

- (34) Noon oughte to *sweare absolutely* for a thinge but that by his owne eyen he be
sure and certeyn that it is soo (Caxton 1489)

Delexicalization or semantic bleaching is a natural result of the grammaticalization process and can account for new meanings. By the end of the 16th century “the modal meaning is competing with a new intensifying sense” (Partington 1993; see (35)).

- (35) That they be very *absolutely skilfull* (1570)

In the present chapter, *absolutely* has been shown to be rising in frequency in recent British English. The increase in frequency is accompanied by syntactic, semantic and pragmatic changes indicating that grammaticalization is still ongoing. Using corpora from two different time periods in present-day English makes it possible to reconstruct the later stages in the development of *absolutely* from an intensifier to an emphasize and to a discourse marker.

An approach to describe how grammaticalization proceeds across different linguistic levels has been proposed by Himmelmann (2004) among others (see also Traugott 2008: 222). According to Himmelmann (2004), grammaticalization is characterized by three types of expansion corresponding to different stages of the grammaticalization process. In the first stage (host-class expansion), the grammatical form will increase its range of collocations resulting in an increase in type frequency. As we have seen, *absolutely* has increased its frequency over time both with regard to type and token frequency (see Section 3). Secondly, the grammaticalizing form can spread to new syntactic contexts (syntactic expansion). It has been shown

that *absolutely* declines in the pre-adjectival position and moves into pre-verbal position (*absolutely* increases in frequency from 6.6% to 12.2%; see Section 4.2). *Absolutely* also becomes more frequent as a discourse marker, indicated by an increase in the frequency of the response marker from 18.4% in 1994 to 29.4% in 2014 (see Section 5).

In the third (and most important) step, *absolutely* spreads to new semantic and pragmatic contexts (semantic-pragmatic expansion). With respect to the semantic-pragmatic expansion of *absolutely*, it has developed in a sequential manner from a maximizing degree-indicating function to an emphaser of truth value. For example, when *absolutely* collocates with verbs such as *leap up* with telic meaning it no longer has a degree meaning. With attitudinal stative predicates such as *love* or *hate*, it is an emphaser closely associated with modality (cf. *definitely*, *really*). *Absolutely* is also an emphaser in other syntactic contexts such as modal auxiliaries or pronouns.

According to Traugott (2008: 235), the semantic changes can be described as possible steps or “fine-grained structural differences” that are constrained by plausible structural shifts along a continuum. The developments can be summarized in the following change-schema:

Degree-modifier > emphaser > discourse marker

Some evidence for this trajectory in the development of *absolutely* is that the different steps can be taken to represent an increase of subjectivity (Athanasidou 2007: 561). *Absolutely* is subjective in the sense that it involves “the speaker’s subjective assessment and evaluation of intensity, position on a scale, ordering of alternatives, etc.” (Traugott 2006: 343). The increase in subjectivity has consequences for both the synchronic and diachronic analysis of *absolutely*. Athanasidou (2007), for example, finds commonalities between *absolutely*’s spread from intensification to emphasis and a rise in subjectivity. Subjectivity also accounts for the further development of *absolutely* into discourse where “speakers now commit themselves to their subjective view of the proposition they are responding to” (Athanasidou 2007: 562).

Moreover, the trajectory from intensifier to emphaser has been found in the diachronic study of other intensifiers. Mendéz-Naya (2008: 283) investigated the development of the adverb *downright* on the trajectory for intensifiers illustrated above. However, in the analysis of *downright*, the ordering was from an emphaser of truth value to intensifier rather than from intensifier to emphaser. Rissanen (2008) explored the grammaticalization of the moderator *rather* on the syntactic path from (an emphasizing) subjunct to modifier of adjectives, and Nevalainen and Rissanen (2002) showed convincingly that the trajectory could describe how *fairly* and *pretty* were used in new syntactic contexts and extended their functional range ultimately gaining adverbial status as intensifying modifier (*pretty*) or as downtoning moderator (*fairly*).

7. Conclusion

The description of the semantic and pragmatic developments over time also requires a broad sociolinguistic analysis, taking into account age and gender. However, the changes described seem to go against the predictions from a sociolinguistic variationist approach. In the majority of linguistic changes, when a linguistic element is incoming, the assumptions are that it will be used mainly by younger speakers and, secondly, that the development to new contexts will be led by female speakers. As far as *absolutely* is concerned, my analysis has shown that younger speakers use it less often than older speakers and that it was first used by male speakers and only later spread to female speakers. We may speculate that older (female) speakers may use *absolutely* consciously and with a higher frequency because of its expressive and hyperbolic qualities (cf. Labov 1990 on changes ‘from above’ operating above the level of conscious awareness). Claridge (2011: 89) found, for example, that women used more hyperbole than men (based on the BNC1994) but that there were differences depending on the individual hyperbolic item. Younger speakers, on the other hand, may prefer other more fashionable intensifiers and they use response marker such as *cool*, *right* or *yeah* rather than *absolutely*.

References

Corpora and software

BNC1994 = *The British National Corpus*. <<http://www.natcorp.ox.ac.uk/>>

BNC2014S = *The British National Corpus 2014*. <<https://cqpweb.lancs.ac.uk/bnc2014spoken/>>

CQP = *Corpus Query Processor*. <<https://cqpweb.lancs.ac.uk/>>

Other references

- Aijmer, Karin. 2018. ‘That’s well bad’: Some new intensifiers in spoken British English. In *Corpus Approaches to Contemporary British Speech. Sociolinguistic Studies of the Spoken BNC2014*, Vaclav Brezina, Robbie Love & Karin Aijmer (eds), 60–95. London: Routledge.
- Altenberg, Bengt. 1991. Amplifier collocation in spoken English. In *English Computer Corpora. Selected Papers and Research Guide*, Stig Johansson & Anna-Brita Stenström (eds), 127–147. Berlin: De Gruyter.
- Athanasiadou, Angeliki. 2007. On the subjectivity of intensifiers. *Language Sciences* 29: 554–565. <https://doi.org/10.1016/j.langsci.2007.01.009>
- Beeching, Kate. 2016. *Pragmatic Markers in British English. Meaning in Social Interaction*. Cambridge: CUP.
- Bolinger, Dwight. 1972. *Degree Words*. The Hague: Mouton.
- Cacchiani, Silvia. 2005. Local vehicles for intensification and involvement: The case of English intensifiers. In *Pragmatics Today*, Piotr Cap (ed.), 401–419. Frankfurt: Peter Lang.

- Carretero, Marta. 2010. 'You're absolutely right!!' A corpus-based contrastive analysis of 'absolutely' in British English and *absolutamente* in Peninsular Spanish with special emphasis on the relationship between degree and certainty. *Languages in Contrast* 10(2): 194–222. <https://doi.org/10.1075/lic.10.2.05scar>
- Claridge, Claudia. 2011. *Hyperbole in English. A Corpus-Based Study of Exaggeration*. Cambridge: CUP.
- Denison, David. 2001. Gradience and linguistic change. In *Historical Linguistics 1999: Selected Papers from the 14th International Conference on Historical Linguistics, Vancouver, 9–13 August 1999*, [Current Issues in Linguistic Theory 215], Laurel J. Brinton (ed.), 119–144. Amsterdam: John Benjamins.
- Fahy, Patrick J. 2002. Use of linguistic qualifiers and intensifiers in a computer conference. *The American Journal of Distance Education* 16(1): 5–22. https://doi.org/10.1207/S15389286AJDE1601_2
- Fuchs, Robert. 2017. Do women (still) use more intensifiers than men? Recent changes in the sociolinguistics of intensifiers in British English. *International Journal of Corpus Linguistics* 22(3): 345–374. <https://doi.org/10.1075/ijcl.22.3.03fuc>
- Hardie, Andrew. 2018. Using the spoken BNC2014 in CQP web. In *Corpus Approaches to Contemporary British Speech. Sociolinguistic Studies of the Spoken BNC2014*, Vaclav Brezina, Robbie Love & Karin Aijmer (eds), 27–30. London: Routledge.
- Hessner, Tanja & Gawlitzek, Ira. 2017. *Totally or slightly different? A Spoken BNC2014-based investigation of female and male usage of intensifiers*. *International Journal of Corpus Linguistics* 22(3): 403–428. <https://doi.org/10.1075/ijcl.22.3.05hes>
- Himmelman, Nikolaus P. 2004. Lexicalization and grammaticalization: Opposite or orthogonal. In *What Makes Grammaticalization?: A Look from its Fringes and Components*, Walter Bisang, Nikolaus P. Himmelman & Björn Wiemer (eds), 21–42. Berlin: Mouton de Gruyter.
- Ito, Rika & Tagliamonte, Sali. 2003. *Well weird, right dodgy, very strange, really cool*: Layering and recycling in English intensifiers. *Language in Society* 32: 257–279. <https://doi.org/10.1017/S0047404503322055>
- Killie, Kristin. 2004. Subjectivity and the English progressive. *English Language and Linguistics* 8(1): 25–45. <https://doi.org/10.1017/S1360674304001236>
- Labov, William. 1990. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change* 2: 205–254. <https://doi.org/10.1017/S0954394500000338>
- Love, Robbie, Dembry, Claire, Hardie, Andrew, Brezina, Vaclav & McEnery, Tony. 2017. The Spoken BNC2014 – designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3): 311–318. <https://doi.org/10.1075/ijcl.22.3.02lov>
- McCarthy, Michael. 2003. Talking Back: 'Small' interactional response tokens in everyday conversation. *Research on Language and Social Interaction* 36(1): 33–63. https://doi.org/10.1207/S15327973RLSI3601_3
- Macaulay, Ronald. 2006. *Pure grammaticalization: The development of a teenage intensifier*. *Language Variation and Change* 18: 267–283. <https://doi.org/10.1017/S0954394506060133>
- Mendéz-Naya, Belen. 2008. On the history of *downright*. *English Language and Linguistics* 12(2): 267–287. <https://doi.org/10.1017/S1360674308002621>
- Murphy, Bróna. 2010. *Corpus and Sociolinguistics. Investigating Age and Gender in Female Talk* [Studies in Corpus Linguistics 38]. Amsterdam: John Benjamins.

- Nevalainen, Terttu & Rissanen, Matti. 2002. *Fairly pretty or pretty fair?* On the development and grammaticalization of English downtoners. *Language Sciences* 24: 359–380.
[https://doi.org/10.1016/S0388-0001\(01\)00038-9](https://doi.org/10.1016/S0388-0001(01)00038-9)
- Núñez-Pertejo, Paloma & Palacios Martínez, Ignacio M. 2014. *That's absolutely crap, totally rubbish:* the use of the intensifiers *absolutely* and *totally* in the spoken language of British adults and teenagers. *Functions of Language* 21(2): 210–237. <https://doi.org/10.1075/fol.21.2.03pal>
- Palacios Martínez, Ignacio M. & Núñez-Pertejo, Paloma. 2012. *He's absolutely massive. It's a super day. Madonna, she is a wicked singer.* Youth language and intensification: A corpus-based study. *Text & Talk* 23(6): 773–796. <https://doi.org/10.1515/text-2012-0036>
- Paradis, Carita. 1997. *Degree Modifiers of Adjectives in Spoken British English*. Lund: Lund University Press.
- Paradis, Carita. 2000. *It's well weird.* Degree modifiers of adjectives revisited: The nineties. In *Corpora Galore. Analysis and Techniques in Describing English*, John Kirk (ed.), 147–160. Amsterdam: Rodopi.
- Paradis, Carita. 2003. Between epistemic modality and degree: The case of *really*. In *Modality in Contemporary English*, Roberta Facchinetti, Frank Palmer & Manfred Krug (eds), 197–220. Berlin: Mouton de Gruyter.
- Partington, Alan. 1993. Corpus evidence of language change. The case the intensifier. In *Text and Technology: In Honour of John Sinclair*, Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds), 177–192. Amsterdam: John Benjamins. <https://doi.org/10.1075/z.64.12par>
- Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rissanen, Matti. 2008. From 'quickly' to 'fairly': On the history of *rather*. *English Language and Linguistics* 12(2): 345–359. <https://doi.org/10.1017/S1360674308002657>
- Stenström, Anna-Brita, Andersen, Gisle & Hasund, Ingrid Kristine. 2002. *Trends in Teenage Talk. Corpus Compilation, Analysis and Findings* [Studies in Corpus Linguistics 8]. Amsterdam: John Benjamins.
- Swales, John & Burke, Amy. 2003. 'It's really fascinating work': Differences in evaluative adjectives across academic registers. In *Corpus Analysis: Language Structure and Language Use*, Pepi Leistyna & Charles F. Meyer (eds), 1–18. Amsterdam: Rodopi.
https://doi.org/10.1163/9789004334410_002
- Tagliamonte, Sali. 2008. So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics* 12(2): 361–394. <https://doi.org/10.1017/S1360674308002669>
- Tagliamonte, Sali A. 2016. *Teen Talk: The Language of Adolescents*. Cambridge: CUP.
- Tagliamonte, Sali & Roberts, Chris. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series *Friends*. *American Speech* 80(3): 280–300.
<https://doi.org/10.1215/00031283-80-3-280>
- Tao, Hongyin. 2007. A corpus-based investigation of *absolutely* and related phenomena in spoken American English. *Journal of English Linguistics* 35(5): 5–29.
<https://doi.org/10.1177/007542406296615>
- Traugott, Elizabeth Closs. 2003. Constructions in grammaticalization. In *The Handbook of Historical Linguistics*, Brian D. Joseph & Richard D. Janda (eds), 624–647. Oxford: Blackwell.
- Traugott, Elizabeth Closs. 2006. The semantic development of scalar focus modifiers. In *The Handbook of the History of English*, Ans van Kemenade & Bettelou Los (eds), 335–359. Malden: Blackwell. <https://doi.org/10.1002/9780470757048.ch14>

Traugott, Elizabeth Closs. 2008. Grammaticalization, constructions and the incremental development of language: Suggestions from the development of degree modifiers in English. In *Variation, Selection, Development: Probing the Evolutionary Model of Language Change*, Regine Eckardt, Gerhard Jäger, & Tonjes Veenstra (eds), 219–250. Berlin: Mouton de Gruyter.

Wright, Susan. 1994. The mystery of the modal progressive. In *Studies in Early Modern English*, Dieter Kastovsky (ed.), 467–485. Berlin: Mouton de Gruyter.

Xiao, Richard & Tao, Hongyin. 2007. A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic Studies* 1(2): 241–273. <https://doi.org/10.1558/sols.v1i2.241>

Appendix 1. Speaker age in BNC1994DS and BNC2014S

BNC1994DS				BNC2014S				
Age	Words	RF	NF	Words	RF	NF	LL	Sig.level
0–14	435,286	20	46.0	69,354	12	173.0	11.20	<0.001
15–24	596,113	89	149.3	962,346	128	133.0	0.70	n.s.
25–34	816,024	80	98.0	385,311	97	251.8	38.75	<0.0001
35–44	825,857	114	138.0	662,217	142	214.4	12.37	<0.0001
45–59	859,736	144	167.5	532,208	160	300.6	25.83	<0.0001
60+	783,594	131	187.7	940,003	394	419.2	94.38	<0.0001

Appendix 2. Cross-tabulation of speaker age and gender in BNC1994DS

BNC1994DS						
Age	Sex	Words	RF	NF	LL	Sig.level
0–14	F	181,726	7	37.3	0.45	n.s.
	M	247,560	13	52.5		
15–24	F	583,136	52	135.7	9.11	<0.01
	M	212,977	37	173.7		
25–34	F	528,041	45	85.2	2.43	n.s.
	M	287,983	35	121.5		
35–44	F	508,501	69	135.7	0.05	n.s
	M	317,356	45	141.8		
45–59	F	538,357	66	122.6	16.67	<0.0001
	M	321,379	78	242.7		
60+	F	480,086	102	212.5	422.98	<0.0001
	M	303,508	29	95.6		

Appendix 3. Cross-tabulation of speaker age and gender in BNC2014S

BNC2014S						
Age	Sex	Words	RF	NF	LL	Sig.level
0–14	F	48,179	11	228.3	3.50	n.s.
	M	21,175	1	47.2		
15–24	F	622,710	49	151.0	1.14	n.s
	M	339,636	34	100.1		
25–34	F	198,767	51	256.6	0.04	n.s
	M	186,544	46	246.6		
35–44	F	410,949	102	248.2	6.01	<0.05
	M	251,268	40	159.2		
45–59	F	334,758	127	379.6	20.34	<0.0001
	M	197,450	33	167.1		
60+	F	343,657	243	707.1	101.94	<0.0001
	M	596,346	151	253.2		

Appendix 4. Adjectives and speaker age

Age	BNC1994DS			BNC2014S			LL	Sig. level
	Words	RF	NF	Words	RF	NF		
0–14	435,286	11	25.3	69,354	3	43.3	0.61	n.s.
15–24	596,113	48	80.5	962,346	47	48.9	5.89	<0.05
25–34	816,024	41	50.2	385,311	31	80.5	3.80	n.s.
35–44	825,857	60	72.7	662,213	31	46.8	4.11	<0.05
45–59	859,736	58	67.5	532,208	51	95.8	5.89	n.s.
60+	783,594	49	62.5	910,003	163	173.5	48.80	<0.0001

Appendix 5. Verbs and speaker age

Age	BNC1994DS			BNC2014S			LL	Sig.level
	Words	RF	NF	Words	RF	NF		
0–14	–	–	–	693,54	5	72.1	–	–
15–24	596,113	18	30.2	962,346	31	32.2	0.05	n.s.
25–34	816,024	16	19.6	385,711	21	54.6	9.50	<0.01
35–44	825,857	17	20.6	662,217	17	25.7	0.41	n.s.
45–59	859,736	17	19.8	532,208	42	78.9	26.29	<0.0001
60+	783,594	13	16.6	94,0003	39	41.5	9.30	<0.0001

Two sides of the same coin?

Tracking the history of the intensifiers *deadly* and *mortal*

Zeltia Blanco-Suárez
University of Cantabria

This chapter presents a corpus-based study on the origin and development of two death-related intensifiers: the adjectives and adverbs *deadly* and *mortal*. The historical sources consulted reveal that these forms have progressively adopted more general meanings, that is, they have come to be grammaticalized as intensifiers over time. Two semantic variables, type of meaning (descriptive, affective, or intensifying, along the lines of Adamson 2000) and semantic prosody (Stubbs 1995), were central to the collocational diachronic analysis undertaken here. The study focuses on British English and covers the history of *deadly* and *mortal* from their origins in Old English and Middle English, respectively, to the 20th century.

1. Introduction

This chapter considers two intensifiers from the semantic field of death: the vernacular *deadly* and its Anglo-Norman counterpart *mortal*. Their respective intensifying functions are illustrated in (1) and (2), and can be paraphrased as *very serious* and *in a great hurry*:

- (1) “You’re pulling my leg!” “Absolutely not. I’m *deadly* serious. What’s the matter?”
(BYU-BNC. 1991. Stephanie Howard. *Battle for Love*)
- (2) ‘And they made off southerly?’ asked Cadfael [...] ‘Southwest,’ amended the young man Martin. ‘And in a *mortal* hurry by all the signs.’
(BYU-BNC. 1993. Ellis Peters. *The Holy Thief*)

In particular, this study sets out to trace the diachronic development of *deadly* and *mortal*, since to the best of my knowledge their history has thus far remained largely unexplored. Although their intensifying uses in Present-day English (PDE)

seem to be somewhat worn out and subject to collocational restrictions,¹ as in the case of *deadly serious* or *mortal hurry* in (1) and (2) above, their degree function as intensifiers is well-documented at previous stages of the language. The diachronic evolution of these death-related forms, therefore, will be the focus of the current study. The development of *deadly* and *mortal*² will be traced here through a collocational³ analysis with evidence from a number of diachronic corpora and electronic collections (cf. Section 3 below), from Old English (OE) to PDE.

The chapter is organized as follows. Section 2 provides an overview of the category of intensifiers and of issues which are of particular concern to the development of this category. Section 3 introduces the resources which have been consulted for the present study. The different uses of *deadly* and *mortal* over time are the subject of Section 4, while Section 5 looks at their grammaticalization. Finally, Section 6 presents some concluding remarks and suggestions for future research.

2. Intensifiers: An overview

Intensifiers are devices, multifarious in nature, which “scale a quality, whether up or down or somewhere between the two” (Bolinger 1972: 17). Intensifiers may range from suprasegmental features such as intonation to lexical or morphosyntactic devices, encompassing, among many others, intensifying prefixes (e.g. *super healthy*), degree adverbs such as *absolutely* or *very*, and forms such as *fucking* or *to be dying to*, which inherently correspond to a superlative in strength. Although the expression of degree has traditionally been associated with adverbs and adjectives are cited as prototypically gradable categories, any part of speech can in fact be intensified, since even non-gradable words can adopt a degree reading in context (cf. Cacchiani 2017: 7–8; Paradis 2008: 322–326). This is the case, for instance, with the adjective *British*. As a rule, it does not accept degree adverbs, and yet we

1. In the *British National Corpus*, for instance, *mortal* occurs with an intensifying meaning in the collocations *mortal hurry*, *mortal embarrassment*, and *mortal importance* (one token each). Intensifying collocations of *deadly* in the same corpus include *deadly serious* (44 tokens), *deadly earnest* (8), and *deadly dull* (7).

2. The words *deadly* and *mortal* can be used as either adjectives or adverbials, but the focus of the chapter is on their use as an intensifier.

3. In the present chapter, *collocation* refers to the association of a node (in this case *deadly* or *mortal*) and a subsequent element (a noun, in the case of the adjectives *deadly* and *mortal*, and an adjective or participial form in the case of the adverbs *deadly* and *mortal*) (cf. Sinclair 1991; Stubbs 1995, among others). *Collocates* are those items which combine with the nodes (*deadly* and *mortal* here).

may easily imagine a context in which we describe someone as being *very British*, alluding perhaps to the dress or particular behavioural traits of the person. On this account, the expression of degree or gradability cannot be restricted to adverbs or any word class in general. In this regard, Paradis (2008) introduces a framework, the Lexical Meaning as Ontologies and Construals model (LOC), which projects a harmonic relationship between the degree modifier and the element which is modified (Paradis 2008: 323). Nonetheless, as in the above case of *British*, occasional mismatches may occur between the modifier and its head.

Based on the type of scale which they invoke, intensifiers or degree modifiers can be classified into totality and scalar. Totality modifiers are bounded and evoke closed scales, as they are conceived in terms of either-or relationships. For example, *dead* applies to referents which can either be dead or alive, admitting only totality degree modifiers (e.g. *completely*). In turn, scalar modifiers are unbounded and signal various degrees on an imaginary degree scale (e.g. *scarcely/quite/tremendously interesting*). Furthermore, in Paradis' (2008) framework, totality modifiers can be further subdivided into maximizers (e.g. *totally*, indicating the maximum/upper end of the degree scale) and approximators (e.g. *almost*), which suggest that the quality itself is less than that maximum/upper end of the scale. Similarly, scalar modifiers can be divided into three different types, depending on the part of the scale which they indicate: boosters denote values within the upper range of the scale (e.g. *very, terribly*), moderators apply to the mid-range of the scale (e.g. *rather*), and diminishers indicate low degree on an imaginary degree scale (e.g. *slightly*). *Deadly* and *mortal*, the forms at issue here, are scalar modifiers, specifically boosters.

The most characteristic feature of intensifiers is probably their tendency to renew and change, which has led them to be at the forefront of academic debate for over a century. Their "fevered invention and competition" (Bolinger 1972: 17) has been discussed from various perspectives, allowing for a wealth of sociolinguistic, diachronic, and synchronic studies that address well-known features of this category, such age- and gender-related uses, as well as their distribution across varieties, registers, and historical periods (cf., among others, Aijmer 2018; Breban & Davidse 2016; Claridge & Kytö 2014; D'Arcy 2015; Fuchs 2017; Macaulay 2006; Méndez-Naya 2006; Méndez-Naya & Pahta 2010; Nevalainen 2008; Nevalainen & Rissanen 2002; Palacios Martínez & Núñez-Pertejo 2012; Peters 1994; Tagliamonte 2008).

Another characteristic of intensifiers concerns their origin. They typically derive from lexical or content words, but develop increasingly grammatical or abstract meanings over time through a process of grammaticalization (Heine 2003; Hopper 1991; Hopper & Traugott 2003; Lehmann 1985). This gradual loss of semantic content, desemanticization or semantic bleaching (Heine 2003) is also concomitant with the expansion of their collocational range (Lorenz 2002; Partington 1993) and semantic prosodies (Stubbs 1995), which is even more noticeable in the case

of originally negative intensifiers. For instance, forms such as *terribly* have come to allow inherently positive collocates (e.g. *good*). In addition to bleaching and grammaticalization, the rise of intensifiers or degree modifiers is often associated with increased subjectivity (cf. Athanasiadou 2007; Traugott 1995, 2003, 2007), that is, with the explicit encoding of the speaker's or writer's point of view, beliefs, and attitudes. Their meanings, as noted by Traugott (1989: 35), are argued to become increasingly based on the speaker's subjective belief state/attitude towards the proposition, a process which is diachronically known as subjectification (cf. López-Couso 2010; Traugott 2010).

The inherently subjective nature of intensifiers was also the focus of Adamson (2000), who suggested a developmental cline of subjectivity for these words, based on evidence for *lovely*: descriptive > affective > intensifying (2000: 55). Thus, *lovely* (< OE *luflic* 'loving, amiable') was originally an adjective which described objective attributes, but adopted more affective or inherently subjective meanings ('physically beautiful') through subjectification in the Middle English (ME) period. Finally, in the 19th century *lovely* acquired intensifying readings ('very'), a change which is associated with syntactic reanalysis: the movement of *lovely* to the left of the adjective string in combinations with other adjectives triggered its reinterpretation as an intensifier.

In what follows, the development of *deadly* and *mortal* will be explored, considering for this purpose two semantic variables: the type of meaning conveyed by these forms and their semantic prosody (Louw 2000; Partington 2015; Smith & Nordquist 2012; Stubbs 1995). For the first variable (type of meaning), Adamson's (2000) model of evolution for intensifiers (descriptive > affective > intensifying) was taken as a reference. The second variable concerns semantic prosody, that is, the collocational preferences of the forms under study.⁴ Since *deadly* and *mortal* are inherently negative, higher rates of occurrence with positive and neutral items could in principle suggest a more grammaticalized status and higher productivity as intensifiers, given that the more widely they collocate, the higher their grammaticalization (Lorenz 2002: 144; Partington 1993: 183). However, before discussing the different findings for these death-related forms in light of their meanings and semantic prosody and their gradual grammaticalization (Sections 4 and 5 below), some remarks on the sources of evidence used here seem to be in order.

4. Unlike some studies on semantic prosody (Smith & Nordquist 2012; Stubbs 1995), in this chapter I have added the category *neutral semantic prosody* for convenience. This term accounts for all those collocations which can be considered as neither positive nor negative. This is the case of *body*, for instance, in contrast to *beauty* and *hatred*, which are inherently positive and negative, respectively. Moreover, those cases in which the form in question may be contextually interpreted (e.g. *proud*), were also counted as neutral.

3. Data sources

Data for the present chapter have been drawn from several corpora and electronic collections. In particular, three diachronic corpora have been selected: the *Helsinki Corpus of English Texts* (HC) for the OE and ME period, the *Early English Books Online Corpus*, version 1.0 (EEBOCorp 1.0) for EModE, and *A Representative Corpus of Historical English Registers*, version 3.2 (ARCHER 3.2) for the 20th-century data. *Eighteenth Century Fiction* (ECF) and *Nineteenth Century Fiction* (NCF) were used to retrieve evidence for the LModE period, specifically for the 18th and 19th centuries.

The HC, EEBOCorp 1.0, and ARCHER are multi-genre corpora, featuring a wide range of text types, such as philosophical, historical, fictional, and legal texts, as well as sermons, diaries, and letters. The main difference between them lies in the periods covered by each corpus, and in their size. Thus, the HC expands from OE to EModE, including texts from roughly 850 to the early 18th century. EEBOCorp 1.0 is targeted at EModE, and ARCHER provides evidence for the period 1650–1990. ARCHER also comprises data for American English, but for the present chapter, only the British component was chosen. Moreover, the EModE section from the HC was disregarded, since data for this period were retrieved from EEBOCorp 1.0. In addition to the different periods, these corpora vary in size: the HC contains approximately 1.5 million words, the British English component of ARCHER 3.2 has circa 2 million words, and EEBOCorp 1.0 has some 525 million words.

Given the limited evidence for *deadly* and *mortal* in ARCHER 3.2., two electronic collections were used to obtain further data for LModE, namely ECF and NCF. These vast resources, with 11 and 37.5 million words, respectively, are produced by Chadwyck-Healey and contain fictional works of well-known authors from the British Isles, including Jonathan Swift, Samuel Richardson, Jane Austen, the Brontës, Mary Shelley, and Charles Dickens, among others. Taken together, these two corpora cover novels from the very beginning of the 18th century to the very beginning of the 20th century (1700–1780 in the case of ECF and 1782–1903 NCF).

The concordance function of WordSmith 6.0 was used to extract all tokens of *deadly* and *mortal* in the HC, EEBOCorp 1.0, and ARCHER 3.2. The wildcard * was used in order to allow for the different orthographic variants of the forms at issue in earlier periods of the language (e.g. *mortel* or *deedli*). In addition, all the irrelevant examples had to be manually discarded, as was the case with the noun *mortal* or with other adjectives and adverbs, leaving only the occurrences which corresponded to the forms in question. Finally, all the relevant tokens were exported to an Excel file for a collocational analysis. By contrast, ECF and NCF cannot be searched for with WordSmith 6.0. In these particular cases, the search interface of these electronic collections was chosen and all the instances were then manually

exported to Excel. Moreover, given that the corpora and electronic collections used here are not tagged for part of speech, it was necessary to determine manually whether each of the occurrences of *mortal* and *deadly* corresponded to the adjectival or the adverbial form.

4. *Deadly* and *mortal*: A historical account

4.1 Preliminary remarks

Before discussing the results from the different sources consulted, it seems appropriate to consult the two historical dictionaries par excellence, the OED and MED, for a broad overview of how *deadly* and *mortal* have changed over time.

According to the OED, *deadly* is a Germanic form dating to OE. At this time, it referred to anything which could lead to death (e.g., a disease or a wound), but it could also apply to humans (e.g. *deadly man*), implying that we are all subject to death. These early meanings of the adjective and adverb would be taken as descriptive or literal on Adamson's (2000) cline. In ME, however, *deadly* adopted non-literal uses, in which death was evoked by alluding to qualities such as darkness, silence, or paleness. In these cases, *deadly* implies greater subjectivity, for it is the speaker/writer who associates these properties with death by describing them as *deadly*, in that they are not in themselves fatal. They would therefore count as affective in Adamson's model, being more subjective than purely descriptive examples. At a final stage, *deadly* developed intensifying readings. The first to develop this function was the adverb, whose earliest intensifying records date from the beginning of the 15th century, in the collocation *deadly dill* 'stupid'. By contrast, the adjective is first documented as an intensifier in the OED in the 17th century, in the collocation *deadly drinker*, which describes a person who drinks immoderately. Therefore, in view of the MED and OED evidence, *deadly* seems to have undergone the stages initially described by Adamson: descriptive > affective > intensifying (2000: 55).

In contrast to *deadly*, *mortal* was borrowed from French in the late 14th century, as suggested by both the OED and the MED. However, unlike *deadly*, which progressively developed affective meanings, *mortal* could occur from its earliest records with both literal and more metaphorical uses. Thus, in addition to collocations like *mortal wounds*, it was found with non-literal collocates such as *hate* or *enemy*, which entail greater subjectivity from the speaker/writer. Finally, in the 18th century the first intensifying uses of the adjective and adverb emerged, in collocations such as *take mortal pains* [to do something] and *to be mortal fond* [of something]).

The following sections present the data from the different diachronic corpora and electronic collections, thus aiming to expand the initial evidence provided by the historical dictionaries and shed further light on the origin and development of *deadly* and *mortal*.

4.2 *Deadly* and *mortal*: The origins

As with the OED, the HC also provides examples of early uses of the adjective *deadly*, with attestations dating as far back as OE. By contrast, only one token of the adverb is recorded in the ME component of the corpus, as a modifier of the participle *idoruen* ('wounded'). Table 1 shows the different collocations of the adjective in OE and ME.

Table 1. The collocations of *deadly* (adj.) in OE and ME
(absolute frequencies and normalized frequencies per million words)

	Collocation	N	pmw
OE	<i>deadly flesh</i>	1	2.42
	<i>deadly lichama</i> ('body')	3	7.26
	<i>deadly life</i>	2	4.84
	<i>deadly lust</i>	1	2.42
	<i>deadly man</i>	1	2.42
	<i>deadly world</i>	1	2.42
	Total	9	21.78
ME	<i>deadly enemy</i>	1	1.64
	<i>deadly flesh</i>	1	1.64
	<i>deadly foe</i>	2	3.29
	<i>deadly life</i>	2	3.29
	<i>deadly man</i>	1	1.64
	<i>deadly sin</i>	10	16.43
	Total	17	27.93

The HC data seem to confirm the initial evidence suggested by the dictionaries. Thus, all the OE tokens of the adjective correspond to descriptive or literal meanings, alluding to the mortal or ephemeral nature of the referent (*flesh*, *man*, *world*, etc.). In ME, however, the adjective appears to have already undergone subjectification, as it can be found with collocates such as *enemy* or *foe*, thereby evincing greater subjectivity (cf. Table 1). Examples (3) to (4) illustrate uses of *deadly* in the HC. In (3), the adverb modifies the participial form *idoruen* ('wounded'), which is descriptive or literal. In turn, the collocation *deadly foe* in (4) is affective, as the adjective cannot be taken literally. By describing an opponent as deadly, the

speaker/writer imposes a subjective conception of reality, given that the enemy is not deadly in and of itself. Nonetheless, it is clear that death can still be invoked in such a context, in that by hating someone one might desire their death. Therefore, *deadly* is considered here affective rather than descriptive.

- (3) þe feondes þe þer weren *deadliche* idoruen. fengen to 3eien.
(CMMARGA. 1150–1250. *Margarete* P92)
- (4) Hwa durste slepen hwil his *deadliche* fa heolde an itohe sweord up on his heaued?
(CMANCRE. 1150–1250. *Ancrene Wisse* P166)

Table 2 shows the different types of meanings for the adjective from OE to ME. Literal or descriptive readings prevail in the data, given that affective *deadly* only represents 17.7% of the ME examples. The adjective is not attested in the corpus with intensifying meanings, which is unsurprising considering that the OED documents the first intensifying uses of *deadly* in the 17th century. The evidence for the adverb in the corpus, by contrast, is very limited, with only one occurrence (Example (3) above).

Table 2. Types of meanings for *deadly* (adj.) in OE and ME

	Descriptive			Affective			Intensifying			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
OE	9	100.0	21.8	0	0.0	0	0	0.0	0	9
ME	14	82.3	23	3	17.7	4.9	0	0.0	0	17

As for semantic prosody (cf. Table 3), in OE the adjective *deadly* is attested as a modifier of neutral heads in almost 89% of the cases (e.g. in collocations such as *deadly world*, *deadly man*, or *deadly life*), although it is also found in combination with negative forms such as *lust*. In turn, in ME negative semantic prosody prevails (with over 76% of the occurrences), although there are also examples of *deadly* with neutral collocates (roughly 23% of the total, as in *deadly life* or *deadly man*). The adverb *deadly* only occurred in the collocation *deadly idoruen* (‘wounded’) in ME, which is clearly negative.

Table 3. Semantic prosody of *deadly* (adj.) in OE and ME

	Negative			Neutral			Positive			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
OE	1	11.1	2.4	8	88.9	19.4	0	0.0	0	9
ME	13	76.5	21.4	4	23.5	6.6	0	0.0	0	17

The HC does not record any instances of *mortal* prior to ME, thus supporting the evidence in the OED that *mortal* was only borrowed during the ME period. Moreover, the HC only provides evidence for the adjective. Its different collocations are given in Table 4.

Table 4. The collocations of *mortal* (adj.) in ME

Collocation	N	pmw
<i>mortal enemy</i>	1	1.64
<i>mortal folk</i>	1	1.64
<i>mortal life</i>	2	3.29
<i>mortal man</i>	3	4.93
<i>mortal nature</i>	1	1.64
<i>mortal thing</i>	2	3.29
<i>mortal violence</i>	1	1.64

All the collocations of *mortal* are considered descriptive, with the exception of *mortal enemy*, which is regarded as affective, along the lines of *deadly enemy* and *deadly foe*. *Mortal enemy* (as in (5)) occurs only once in the HC and is applied to Satan. In this case, the writer is imposing a subjective assessment of the situation, since Satan, unlike a disease or a wound, is not deadly. Nonetheless, the connection with death is very clear, as in the Bible the Devil is regarded as a fallen angel who entices men into sin, hence causing their spiritual death. In all the other collocations *mortal* refers to humans as subject to death, or to referents which are also ephemeral, as in *mortal nature*, *mortal life*, and *mortal man* (see (6)).

- (5) For yt hath dyssoluyde mankynde from þe bytter bonde Of þe *mortall* enmye,
þat vemynousse serpente, (CMMANKIN. 1420–1500. P155)
- (6) god gaue hym wysdom aboue all *mortall* men
(CMFITZJA. 1420–1500. PB5R)

The distribution of the different types of readings for the adjective *mortal* is shown in Table 5. As happened with *deadly*, its Germanic counterpart, *mortal* is also found chiefly with descriptive or literal uses, whereas affective *mortal* accounts for just 10% of the total. Intensifying *mortal* is not yet attested, as this function arose in the 18th century, according to the OED.

Table 5. Types of meanings for *mortal* (adj.) in ME

Descriptive			Affective			Intensifying			Total
N	%	pmw	N	%	pmw	N	%	pmw	
10	90.9	16.4	1	9.1	1.6	0	0.0	0	11

The collocational preferences of *mortal* in ME (cf. Table 6 below) differ considerably from those of *deadly*, since *mortal* combines with neutral forms in over 80% of the occurrences in this period (vs. roughly 24% in the case of *deadly*). Examples of *mortal* with negative polarity items are also found, but positive collocations are absent from the data, as was also the case with *deadly*.

Table 6. Semantic prosody of *mortal* (adj.) in ME

Negative			Neutral			Positive			Total
N	%	pmw	N	%	pmw	N	%	pmw	
2	18.2	3.29	9	81.8	14.8	0	0.0	0	11

Having presented the data for the early history of *deadly* and *mortal*, comprising the OE and ME periods, the following section turns to data from EModE to look at any potential advances in the evolution of these forms.

4.3 *Deadly* and *mortal*: EModE

In comparison with the HC, EEBOCorp 1.0 provides a substantial body of evidence for both *deadly* and *mortal*. The only exception is the zero-adverb *mortal*, for which there are only four tokens. The adjective *deadly* occurs a total of 8,062 times in the corpus, in 485 different collocations. The adverb is considerably less frequent: 920 tokens in 131 different collocations. The most common collocations for the adjective and the adverb in EEBOCorp 1.0 are shown in Table 7. For convenience purposes, it only includes those collocations which are attested at least 100 times in the case of the adjective, and with a minimum of 10 occurrences in the case of the adverb.

The EModE data reveal further changes for *deadly* with respect to the earliest records, most conspicuously in the case of the adverbial form. Despite the overall increase in absolute frequencies, the adjective is still predominantly found with descriptive meanings in EModE, as shown in Table 8. Affective readings amount to almost 28%, and intensifying uses of the adjective are not recorded in the corpus. It should be recalled, however, that the OED dates these intensifying uses to the second half of the 17th century and, therefore, early examples of intensifying *deadly* could have been recorded in EEBOCorp 1.0. By contrast, the adverb is attested with intensifying meanings in the corpus, even though they are not the most common. Like its adjectival counterpart, descriptive uses are the most frequent (approximately 60%), followed by affective ones (36%).

Table 7. The collocations of *deadly* (adj. and adv.) in EModE

	Collocation	N	pmw
<i>deadly</i> (adj.)	<i>deadly blow</i>	252	0.48
	<i>deadly disease</i>	134	0.26
	<i>deadly enemy</i>	775	1.47
	<i>deadly feud</i>	212	1.40
	<i>deadly foe</i>	176	0.33
	<i>deadly hate</i>	114	0.22
	<i>deadly hatred</i>	274	0.52
	<i>deadly poison</i>	659	1.25
	<i>deadly sin</i>	1,452	2.76
	<i>deadly thing</i>	109	0.21
	<i>deadly war</i>	100	0.19
	<i>deadly wound</i>	818	1.56
<i>deadly</i> (adv.)	<i>deadly hate</i>	224	0.40
	<i>deadly sick</i>	74	0.10
	<i>deadly sin</i>	129	0.25
	<i>deadly sting</i>	11	0.00
	<i>deadly wound</i>	41	0.10
	<i>deadly wounded</i>	136	0.26

Table 8. Types of meanings for *deadly* (adj. and adv.) in EModE

	Descriptive			Affective			Intensifying			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
<i>deadly</i> (adj.)	5,821	72.2	11.1	2,241	27.8	4.3	0	0.0	0	8,062
<i>deadly</i> (adv.)	555	60.3	1.1	331	36.0	0.6	34	3.7	0.06	920

Examples (7) to (10) provide a selection of the different types of meanings of the adjective and adverb in EEBOCorp 1.0. *Deadly wounded* in (7) illustrates a descriptive or literal meaning of the adverb; a *deadly fright* in (8) is affective, since in this case the character is scared and shocked, probably fearing for his life. Since the situation or context is not fatal but death can be easily evoked, a descriptive reading is discarded and *deadly* is instead taken as affective. Similarly, (9) reports a situation of religious persecution by Anabaptists on account of their jealousy. Given that an implicit intention of hurting others is present, this example is counted as affective. However, *deadly envy* here could also contextually allow for a degree interpretation: ‘pure/extreme envy’. In turn, Examples (10) to (11) are intensifying: ‘very lazy’ and ‘very proud’, respectively.

- (7) and yet is not wholly rooted vp, as you falsely make him speake, but wounded in the roote so *deadly*, [...] (D00000998555290000. 1606. Anthony Wotton.
A Defence of M. Perkins Booke)
- (8) The last seeing so bloody an execution, [...] hee ran cross the wood in a *deadly* fright. (D00000111470190000. 1658. Gaultier de Coste.
Hymen's Praeludia, or, Loves Master-piece)
- (9) we cannot herein but take notice of the persecuting Spirit, implacable and *deadly* Envy that is in some of these Anabaptists, tending to Persecution in the highest. (D00000162047460000. 1674. William Penn.
The Christian-Quaker and his Divine Testimony)
- (10) Misoponos was put to a Trade, but a *deadly* lazy Boy he was, very dull at learning, and hardly brought to settle to any thing, or stay with any Master. (D00000112731160000. 1669. Richard Brathwaite.
The History of Moderation)
- (11) Yet He's [...] grown so *deadly* proud, since He was made Governour of the Town, [...] (D00000119240840000. Sir Robert Stapylton.
The Tragedie of Hero and Leander)

As was the case for the findings for semantic prosody in the ME period, the adjective and adverb *deadly* show a clear preference for negative items: in almost 87% and 95% of the cases, respectively. Unlike in the ME data, however, *deadly* is now attested as a modifier of positive heads, although to a marginal extent, and also combines with neutral elements (cf. Table 9).

Table 9. Semantic prosody of *deadly* (adj. and adv.) in EModE

	Negative			Neutral			Positive			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
<i>deadly</i> (adj.)	7,007	86.9	13.3	1,047	13.0	2.0	8	0.1	0.02	8,062
<i>deadly</i> (adv.)	872	94.8	1.7	43	4.7	0.1	5	0.5	0.01	920

As in the case of *deadly*, the data for *mortal* in EEBOCorp 1.0 is also considerably expanded, with some 20,697 occurrences of the adjective in 728 different collocations. The adverb, in turn, must have been a rather infrequent form, since there are only four attestations in the corpus: *mortal wounded* (2 tokens), *mortal pale* (1 token), and *mortal sick* (1 token). The most frequent collocations for the adjective *mortal* are provided in Table 10. Given the vast number of examples and collocations, only those combinations of *deadly* + noun which occur at least 100 times in the corpus are provided.

Table 10. The collocations of *mortal* (adj.) in EModE

Collocation	N	pmw
<i>mortal battle</i>	121	0.22
<i>mortal blow</i>	165	0.31
<i>mortal body</i>	1,554	2.96
<i>mortal creature</i>	561	1.07
<i>mortal disease</i>	226	0.43
<i>mortal enemy</i>	1,799	3.42
<i>mortal eye</i>	592	1.13
<i>mortal flesh</i>	247	0.47
<i>mortal foe</i>	340	0.65
<i>mortal hate</i>	115	0.22
<i>mortal hatred</i>	269	0.51
<i>mortal life</i>	1,255	2.39
<i>mortal man</i>	4,567	8.70
<i>mortal nature</i>	151	0.29
<i>mortal part</i>	102	0.19
<i>mortal sickness</i>	109	0.21
<i>mortal sin</i>	2,153	4.10
<i>mortal state</i>	244	0.46
<i>mortal thing</i>	381	0.73
<i>mortal war</i>	241	0.46
<i>mortal wight</i>	237	0.45
<i>mortal wound</i>	693	1.32

Like *deadly*, the adjective *mortal* is most frequent with descriptive readings (85% of the total). Affective uses account for 15% of the occurrences and intensifying *mortal* is not yet attested in EModE (cf. Table 11), which come as no surprise, considering that the OED first documents these uses in the 18th century.

Table 11. Types of meanings for *mortal* (adj.) in EModE

Descriptive			Affective			Intensifying			Total
N	%	pmw	N	%	pmw	N	%	pmw	
17,604	85.1	33.5	3,093	14.9	5.9	0	0.0	0	20,697

Examples (12) and (13) illustrate the uses of *mortal* in the corpus: *mortal folk* is descriptive, alluding to humans as subject to death, and *mortal enemy* is affective, as was the case with *deadly foe* in Example (4).

- (12) O ye *mortal* folke what seke ye then blisfulnes out your self.
(D00000998422230000.1478. Boecius. *De Consolacione Philosophie*)
- (13) beholde and see me I am Iupiter the *mortal* enemye of alle thy lignage
(D00000998449480000. 1473. Raoul Lefèvre. *The Historyes of Troye*)

In terms of semantic prosody (cf. Table 12), *mortal* is most commonly attested as a modifier of neutral elements (in 60% of cases), while *deadly* shows a clear preference for negative heads. As found with *deadly*, combinations of *mortal* with positive heads are also rare.

Table 12. Semantic prosody of *mortal* (adj.) in EModE

Negative			Neutral			Positive			Total
N	%	pmw	N	%	pmw	N	%	pmw	
8,091	39.1	15.4	12,335	59.6	23.5	271	1.3	0.5	20,697

The adverb *mortal*, as noted above, was marginal in the corpus, with only four tokens: *mortal pale* (1), *mortal wounded* (2), and *mortal sick* (1). *Mortal sick* and *mortal wounded* illustrate descriptive or literal meanings, but *mortal pale* shows greater subjectivity and is, rather, affective. All of these show negative semantic prosody.

The following section examines the uses of *deadly* and *mortal* in LModE and PDE, in order to account for any potential change in their types of meaning and semantic prosody over the course of these periods.

4.4 *Deadly* and *mortal*: LModE and PDE

Although the absolute figures for *deadly* and *mortal* in ECF, NCF, and ARCHER decrease considerably in comparison to EEBOCorp 1.0, these sources also provide very valuable insights into the development of the forms. In fact, LModE seems to be a turning point for their intensifying uses. The overall number of collocations in ECF and NCF is relatively high, given that many of them occur just once or twice. Thus, the adjective *deadly* is recorded with 33 different collocates in ECF and with 192 in NCF. Given the impracticality of displaying all these collocations, Table 13 sets out only those combinations of the adjective and adverb which occur twice in these collections. The data from ARCHER are not represented in the table, since the adjective *deadly* is not attested in the 20th century materials from the corpus and its adverbial counterpart is only found in one example, corresponding to the collocation *deadly weary*.

Table 13. The collocations of *deadly* (adj. and adv.) in LModE

		Collocation	N	pmw
ECF	<i>deadly</i> (adj.)	<i>deadly composition</i>	2	0.2
		<i>deadly oath</i>	2	0.2
		<i>deadly passion</i>	2	0.2
		<i>deadly poison</i>	2	0.2
		<i>deadly sin</i>	2	0.2
	<i>deadly</i> (adv.)	<i>deadly fierce</i>	2	0.2
		<i>deadly sore</i>	2	0.2
		<i>deadly strong</i>	2	0.2
NCF	<i>deadly</i> (adj.)	<i>deadly blow</i>	13	0.4
		<i>deadly enemy</i>	14	0.4
		<i>deadly fear</i>	10	0.3
		<i>deadly feud</i>	21	0.6
		<i>deadly hatred</i>	16	0.4
		<i>deadly paleness</i>	14	0.4
		<i>deadly poison</i>	10	0.3
		<i>deadly sickness</i>	10	0.3
		<i>deadly sin</i>	27	0.7
		<i>deadly wound</i>	11	0.3
	<i>deadly</i> (adv.)	<i>deadly cold</i>	5	0.1
		<i>deadly pale</i>	106	2.8
		<i>deadly sick</i>	5	0.1
		<i>deadly white</i>	14	0.4

As far as the types of meanings are concerned (cf. Table 14), the picture for LModE changes considerably with respect to EModE, in that in LModE over half of the occurrences of the adjective *deadly* correspond to affective uses, followed by descriptive ones, and, most importantly, intensifying readings of the adjective are first attested, although they are infrequent. Likewise, the evidence for LModE reveals important changes in relation to the adverb. Whereas descriptive occurrences of the adverb accounted for 60% of the total in EModE, in the 18th-century data intensifying uses are the most frequent (63%), followed by affective and descriptive ones. In the 19th century, by contrast, intensifying *deadly* reduces considerably (to 5%), and instead affective readings represent the vast majority (91%). Descriptive meanings are also infrequent (4%).

Table 14. Types of meanings for *deadly* (adj. and adv.) in LModE

<i>deadly</i> (adj.)	Descriptive			Affective			Intensifying			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
ECF	14	35.9	1.4	22	56.4	2.3	3	7.7	0.3	39
NCF	215	44.9	5.7	242	50.5	6.4	22	4.6	0.6	479

<i>deadly</i> (adv.)	Descriptive			Affective			Intensifying			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
ECF	1	12.5	0.1	2	25.0	0.2	5	62.5	0.5	8
NCF	6	3.9	0.2	139	90.9	3.7	8	5.2	0.2	153

Some illustrative examples of the adjective and adverb from the period are given in (14) to (17). The *deadly potion* in (14) refers to a toxic mixture and is therefore descriptive; (15) implies negative affect or a subjective perception, whereby the narrator refers to the character as being as pale as a corpse, hence being an affective reading of the adverb. Examples (16) and (17), in turn, are intensifying, for in this case *deadly* only admits a degree reading: ‘very strong resolution’ and ‘very curious’, respectively.

- (14) Dalinda failed not to execute her Mistress’s Orders, and having mixed the *deadly* Potion, left the Bowl upon a Table in the next Room,
(ECF.1739. Penelope Aubin. *The Noble Slaves*)
- (15) He was *deadly* pale, and his hair, which hung in dishevelled locks over his face, was clotted with blood.
(NCF. 1858. Robert Michael Ballantyne. *The Coral Island*)
- (16) With this *deadly* Determination, I went Post, from Paris to Flanders, and traced the Colonel from Place to Place, till I found him in a Village, on the Road to Amsterdam.
(ECF. 1765–1770. Henry Brooke. *The Fool of Quality*)
- (17) I’m *deadly* curious to know what keeps you from coming to a finish.
(NCF. 1814. Fanny Burney. *The Wanderer*)

In terms of semantic prosody (cf. Table 15), *deadly* is clearly disfavoured with positive collocates, as also happened in the previous periods. In LModE, however, both the adjective and adverb *deadly* licence more neutral nouns. In fact, the adverb co-occurs with neutral forms in over 93% of the 19th-century data, and in about 50% of the cases in the previous century. The adjective, by contrast, shows a clear bias towards negative collocates. The data for *deadly* in the 20th century, as indicated above, are extremely limited: only one occurrence of the adverb, *deadly weary*, which is inherently negative.

Table 15. Semantic prosody of *deadly* (adj. and adv.) in LModE

<i>deadly</i> (adj.)	Negative			Neutral			Positive			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
ECF	22	56.4	2.3	14	35.9	1.4	3	7.7	0.3	39
NCF	311	64.9	8.3	163	34.0	4.34	5	1.0	0.1	479

<i>deadly</i> (adv.)	Negative			Neutral			Positive			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
ECF	3	37.5	0.3	4	50.0	0.4	1	12.5	0.1	8
NCF	9	5.9	0.2	143	93.5	3.8	1	0.7	0.03	153

In line with the data for *deadly* in LModE and PDE, the number of tokens of *mortal* in ECF, NCF, and ARCHER are drastically reduced: from 20,697 attestations of the adjective in EModE, there are 1,295 occurrences in total in the 18th and 19th centuries. Despite this drop in absolute figures, LModE witnesses important changes in the development of *mortal*, since it is during this period that the first intensifying uses are recorded in the data. Moreover, ECF and NCF expand the evidence for the adverb, which was fairly limited in EBOCORP 1.0 (only four tokens), although there are only 48 occurrences in ECF and NCF altogether. The findings from ARCHER, in line with those for *deadly*, are very restricted: there are only four occurrences of the adjective, in the collocations *mortal man* (two tokens), *mortal enemy*, and *mortal sin* (one token each), and there is just one example of the adverb in the 20th-century data from the corpus, the collocation *mortal hard* (Example (21) below). In the LModE data, the adjective *mortal* is found in 106 different collocations in ECF, and in 342 in NCF. The adverb appears in five collocations in ECF, in combination with the adjectives *dear*, *good*, *ill*, *jealous*, and *sick*, and in 32 different collocations in NCF. Table 16 shows the most frequent collocations of *mortal* in the data.

As in EModE, the most frequent meanings of the adjective *mortal* in LModE are descriptive ones, accounting for about 75% and 80% in the 18th and 19th centuries. Affective *mortal* is the second most frequent meaning (25% and 17%, respectively) and, interestingly, intensifying readings emerge in the 19th century, although they represent less than 3% of the total. In contrast, affective and intensifying meanings of the adverb constitute about 60% of the total in the 18th century, and in the 19th century the intensifying function of the adverb prevails in the data (68%), followed by affective (27%) and descriptive (5%) ones. The distribution of the different types of meanings for the adjective and adverb is provided in Table 17 below.

Table 16. The collocations of *mortal* (adj. and adv.) in LModE

		Collocation	N	pmw
ECF	<i>mortal</i> (adj.)	<i>mortal enemy</i>	17	1.75
		<i>mortal life</i>	16	1.65
		<i>mortal man</i>	37	3.81
		<i>mortal offence</i>	11	1.13
		<i>mortal wound</i>	15	1.55
	<i>mortal</i> (adv.)	<i>mortal ill</i>	2	0.2
		<i>mortal jealous</i>	2	0.2
NCF	<i>mortal</i> (adj.)	<i>mortal agony</i>	17	0.45
		<i>mortal blow</i>	15	0.40
		<i>mortal combat</i>	17	0.45
		<i>mortal creature</i>	17	0.45
		<i>mortal disease</i>	19	0.51
		<i>mortal enemy</i>	37	0.98
		<i>mortal eye</i>	29	0.77
		<i>mortal fear</i>	20	0.53
		<i>mortal life</i>	28	0.74
		<i>mortal man</i>	96	2.55
		<i>mortal offence</i>	17	0.45
		<i>mortal remains</i>	17	0.45
		<i>mortal woman</i>	22	0.59
		<i>mortal wound</i>	18	0.48
	<i>mortal</i> (adv.)	<i>mortal angry</i>	3	0.1
		<i>mortal cold</i>	3	0.1
		<i>mortal hard</i>	2	0.05
		<i>mortal high</i>	2	0.05
		<i>mortal hungry</i>	2	0.05
		<i>mortal long</i>	2	0.05
		<i>mortal still</i>	2	0.05

Examples (18) to (21) include a selection of the different readings of *mortal* in the data. Thus, in (18) it refers to human ears and is hence descriptive; *mortal white* in (19) is like *deadly pale* in (15). In this context death can still be evoked, as the paleness of the character resembles that of a dead body and, therefore, is considered affective. Conversely, in (20) and (21) *mortal* is an intensifier: ‘a strong distaste’ and ‘very hard’.

- (18) Sounds made such a Complication of Horrors as no *mortal* Ears could abide, nor mortal Courage support.
(ECF. 1765–70. Henry Brooke. *The Fool of Quality*)
- (19) The sun was getting up, and *mortal* white he looked about the cutwater.
(NCF. 1883. Robert L. Stevenson. *Treasure Island*)

- (20) Cos they has a *mortal* awersion to washing anythin, I suppose, Sir, replied Mr. Weller.

(NCF. 1837. Charles Dickens. *The Posthumus Papers of the Pickwick Club*)

- (21) It's *mortal* 'ard, sir, that a man o' my years should be tole, 'e in liquor twice in one evenin'!

(ARCHER. 1911besi_d7b. 1911. Rudolf Besier. *Lady Patricia Clare*)

Table 17. Types of meanings for *mortal* (adj. and adv.) in LModE

<i>mortal</i> (adj.)	Descriptive			Affective			Intensifying			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
ECF	214	74.8	22.1	72	25.3	7.4	0	0.00	0	286
NCF	808	80.1	21.5	175	17.3	4.7	26	2.6	0.7	1,009

<i>mortal</i> (adv.)	Descriptive			Affective			Intensifying			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
ECF	3	42.8	0.3	2	28.6	0.2	2	28.6	0.2	7
NCF	2	4.9	0.05	11	26.8	0.3	28	68.3	0.7	41

In the 18th-century data the adjective *mortal* is slightly more frequent in combinations with neutral elements (about 49%), but negative collocates are also common (46%). *Mortal* as a modifier of positive heads is sporadically attested (5%). In the 19th century, however, the percentage of neutral collocates increases to almost 60%, while negative heads represent 39% and positive elements are still rare (3%). As with its adjectival counterpart, the adverb *mortal* selects mostly negative and neutral nouns in ECF and NCF. Nonetheless, positive collocates are also attested. Moreover, unlike the adjective *mortal*, the combinations of the adverb with positive lexemes correlate in all cases with intensifying meanings (cf. Table 18).

Table 18. Semantic prosody for *mortal* (adj. and adv.) in LModE

<i>mortal</i> (adj.)	Negative			Neutral			Positive			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
ECF	132	46.2	13.6	139	48.6	14.3	15	5.2	1.6	286
NCF	394	39	10.5	584	57.9	15.5	31	3.1	0.8	1,009

<i>mortal</i> (adv.)	Negative			Neutral			Positive			Total
	N	%	pmw	N	%	pmw	N	%	pmw	
ECF	5	71.4	0.5	0	0.0	0	2	28.6	0.2	7
NCF	17	41.5	0.5	17	41.5	0.5	7	17	0.2	41

5. The grammaticalization of *deadly* and *mortal*

The two variables examined in the data, type of meaning (descriptive, affective, and intensifying) and semantic prosody (negative, neutral, and positive), have proved to be of crucial importance in accounting for the evolution of *deadly* and *mortal*. This section presents a diachronic overview of the semantic development of these death-related intensifiers, focusing on their eventual grammaticalization.

As shown in Figures 1 and 2⁵, until EModE *deadly* is very closely attached to its original descriptive or literal meaning, ‘causing death’. Affective readings, although considerably less frequent than descriptive ones, are also attested without interruption from ME. The intensifying function of *deadly* is marginal, with scant examples in the 18th and 19th centuries, these in collocations such as *deadly accuracy*, *deadly determination* (Example (16) above), *deadly strong*, and *deadly comfortable*. Moreover, in terms of their collocational preferences (cf. Figures 3 and 4), *deadly* manifests a clear tendency to co-occur with negative polarity items from ME onwards, which is most conspicuous in the case of the adjective, since it combined predominantly with neutral forms in the OE data. Positive collocations and intensifying meanings are clearly disfavoured and only appear occasionally from EModE onwards.

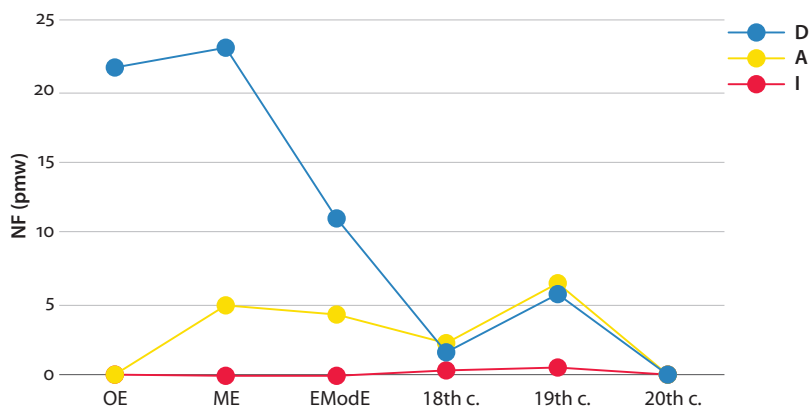


Figure 1. Types of meanings for *deadly* (adj.) over time

5. The following abbreviations are used in the different figures throughout this section: D (descriptive), A (affective), and I (intensifying) as for the types of meanings, and neg (negative), neu (neutral), and pos (positive) as regards semantic prosody.

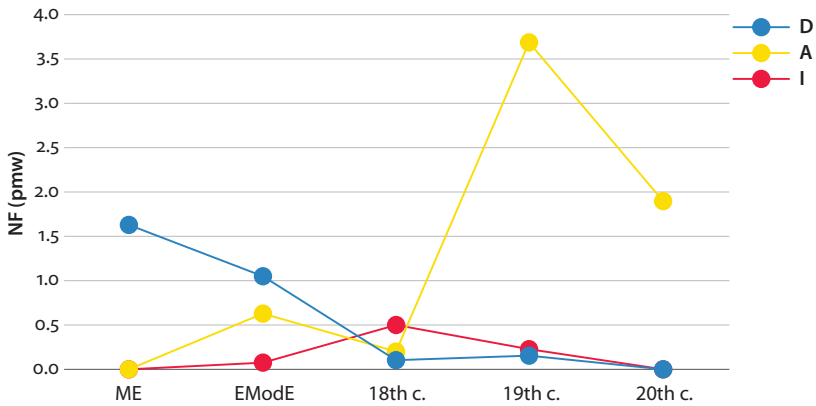


Figure 2. Types of meanings for *deadly* (adv.) over time

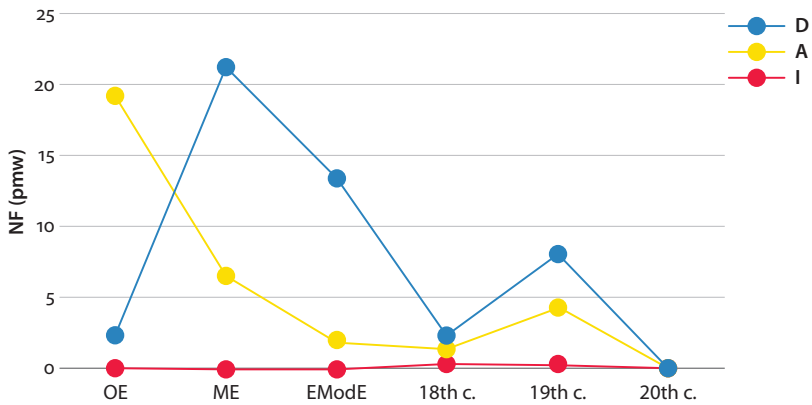


Figure 3. Semantic prosody of *deadly* (adj.) over time

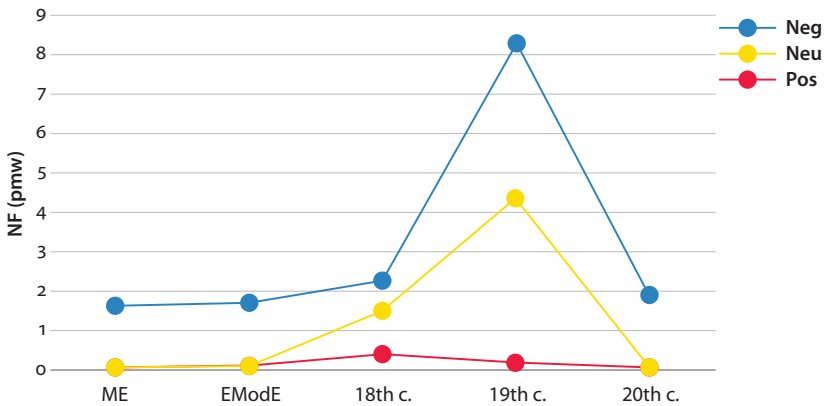


Figure 4. Semantic prosody of *deadly* (adv.) over time

In view of the data, *deadly* illustrates some features which are traditionally related to grammaticalization. Thus, from ME onwards the adjective and adverb occur with increasingly subjective uses as a result of a process of subjectification by collocations such as *deadly foe* or *deadly pale*. That is, they have progressively come to be based more on the speaker's/writer's subjective opinion towards the proposition (Traugott 1995: 31). It is, in fact, the speaker/writer who attributes specific death-related characteristics to a given context (e.g. paleness, silence, etc.): 'characteristic of death' > 'to a degree that is suggestive of death'. Moreover, the adjective and adverb have acquired more abstract or grammatical meanings over time, that is, they have been semantically bleached, hence collocations such as *deadly determination* or *deadly comfortable*, in which *deadly* only admits a degree reading. Their semantic bleaching or desemanticization (Heine 2003: 579) also allows *deadly* to occur in combination with positive items. Nonetheless, positive semantic prosody in the case of the adjective does not always correlate with an intensifying function, since in EModE *deadly* modifies positive nouns such as *pleasure* and *charms*. These, rather, are descriptive collocations, referring to their mortal or ephemeral nature. These findings, therefore, contrast with those of Partington for the "‘terribly’ group of words" (1993: 183–184), whereby intensifiers such as *terribly* or *awfully* allowed increasingly neutral and positive collocates over time, hence expanding their collocational range. In turn, the data for the adverb *deadly* do show a correlation between positive prosody and intensifying meaning, among others in collocations such as *deadly handsome* and *deadly comfortable*.

The diachronic development of *deadly* thus fits into the "modal-to-intensifier shift" claimed by Partington (1993: 181), which is also in keeping with another characteristic of grammaticalization, namely decategorialization (Hopper & Traugott 2003: 106–115). Thus, from major categories such as adjectives or (manner) adverbs, *deadly* has come to be reinterpreted as an intensifier, hence fitting into the grammaticalization cline major > minor category (Hopper & Traugott 2003: 107). In the case of the evolution of the adverb there has also been fixation and scope reduction (Lehmann 1985: 307–308). Originally an adjunct at clausal level (cf. Example (7) above, in which *deadly* modifies the participle *wounded*), the scope of *deadly* was reduced to phrasal level. When having clausal scope, *deadly* could modify verbal or participial forms and allowed for intervening material (e.g. *in the roote so* in Example (7)). As an intensifier, however, *deadly* specialized as a modifier of adjectives and participles and there are no longer any elements between the node and its collocates (e.g. in (17) earlier).

The analysis of the types of meanings and semantic prosody of *deadly* has revealed a further characteristic of grammaticalization, namely persistence (Hopper 1991: 28–30). Thus, despite the occasional uses of *deadly* with positive elements, it occurs recurrently with negative collocates, in line with its primitive meaning.

Descriptive readings are also the most frequent, and it is only in LModE that affective senses become more prominent than descriptive ones. Moreover, the coexistence of descriptive, affective, and intensifying meanings in the 18th and 19th centuries constitutes a case of layering (Hopper 1991: 22–24) or “persistence of older forms and meanings alongside newer forms and meanings” (Hopper & Traugott 2003: 124). The emergence of new intensifying readings implies yet another concomitant of grammaticalization, in particular extension (Heine 2003: 579; Himmelmann 2004: 31–34), since the newly grammaticalized form can be used in additional contexts.

On the same lines as *deadly*, the adjective *mortal* is most frequently used with descriptive senses, whereas affective readings clearly lag behind the descriptive. Intensifying uses of *mortal* are scarce in the data, with few attestations in the 19th century (e.g. *mortal hurry*, *mortal repugnance*, or *mortal disappointment*) (cf. Figure 5). Its adverbial counterpart is relatively infrequent in the data, amounting to only 53 tokens. Unlike the adjective, however, the frequencies for descriptive and affective uses of the adverb are fairly similar in the 18th century. In fact, in the 19th century, intensifying *mortal* becomes the predominant function, followed by affective and descriptive occurrences. In the 20th-century data from ARCHER, the adverb only occurs once, in the collocation *mortal hard*, which is intensifying (cf. Figure 6). The collocational profile of *mortal* shows a similar distribution of the adjective with negative and neutral polarity items. The sharpest contrast is found in ME and EModE, when it is clearly preferred with neutral rather than with negative nouns. Positive nouns, in turn, are not so frequently attracted to combine with the node *mortal* (cf. Figure 7). The differences with regard to semantic prosody are not so marked for the adverb *mortal* (cf. Figure 8), which is only preferred with negative collocates in the 18th-century source materials. In the 19th century, negative and neutral collocates of the adverb are distributed alike, and combinations of *mortal* with positive collocates are only attested in LModE.

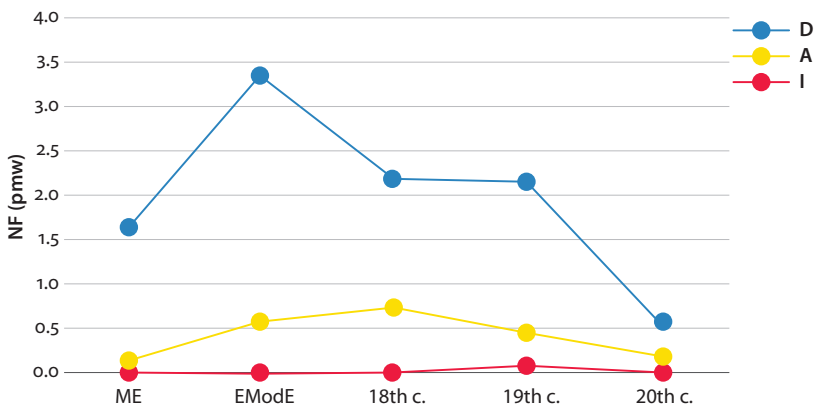


Figure 5. Types of meanings for *mortal* (adj.) over time

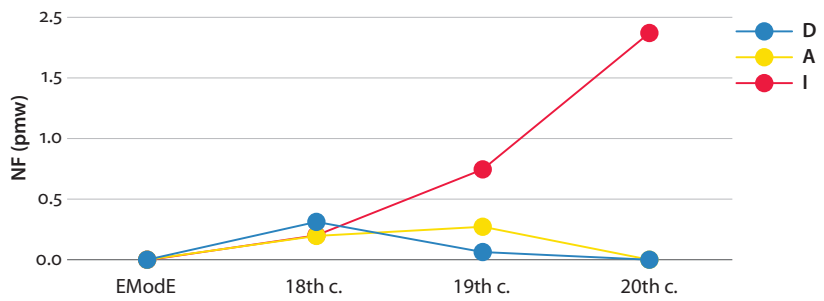


Figure 6. Types of meanings for *mortal* (adv.) over time

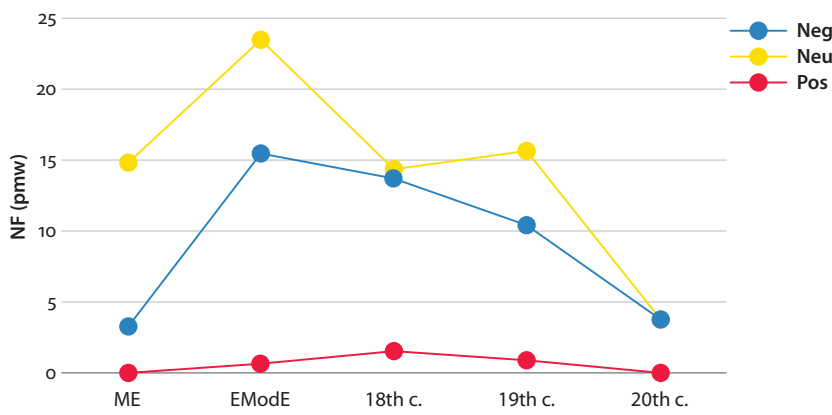


Figure 7. Semantic prosody of *mortal* (adj.) over time

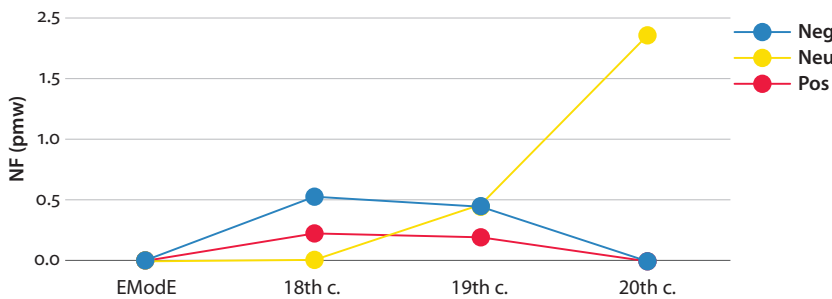


Figure 8. Semantic prosody of *mortal* (adv.) over time

The diachronic evolution of *mortal* also shows several features compatible with grammaticalization. As was the case with its Germanic counterpart *deadly*, *mortal* has been semantically bleached. Intensifying collocations such as *mortal certainty*, *mortal aversion*, *mortal good*, or *mortal cunning* clearly attest to the grammaticalized status of *mortal* and reveal a direct correlation between delexicalization and collocational expansion. Like *deadly*, the development of a new degree or intensifying function suggests decategorialization, that is, the change from a major into a minor category (adjective and manner adverb > intensifier). Moreover, *mortal* has increased in subjectivity over time through a process of subjectification, with affective uses reaching their peak in the 18th century in the case of the adjective and in the 19th century with the adverb. Nonetheless, unlike *deadly*, which was attested first with literal examples in OE and which only developed affective uses in ME, *mortal* was found in the ME data with both affective and descriptive readings. Therefore, the initial stage descriptive > affective in Adamson's cline (2000: 55) cannot be confirmed from the data consulted. The collocational analysis conducted here has also revealed a clear preference of *mortal* for neutral and negative collocates, whereas positive heads are disfavoured. As in the case of *deadly*, this is due to the persistence of the old meanings of *mortal*, a phenomenon which is consistent with grammaticalization. Furthermore, despite the rise of a new intensifying function, the original meanings coexist synchronically with the new ones, which indicates layering (Hopper 1991: 22–24). Unlike the adverb *deadly*, the scant evidence for its counterpart *mortal* in the earliest materials, with only four tokens in EEBORP 1.0, cannot serve to confirm the interplay of fixation and scope reduction.

In summary, despite being close synonyms and appearing to be two sides of the same coin, the development of *deadly* and *mortal* is not quite the same. Whereas *deadly* seems to be more biased than *mortal* towards negative polarity items, *mortal* licences more neutral and even positive collocates. In fact, positive prosodies correlate with intensifying meanings in the case of the adverb *mortal*, which despite its low frequency, also appears to be associated with an intensifying function in the most recent data.

6. Concluding remarks

This chapter has examined the history of *deadly* and *mortal*, from their original literal meaning, applying to referents which could cause death or were subject to death, to their final semantic bleaching and subsequent development as intensifiers ('very'). The evolution of these forms was traced here through a collocational analysis using two variables, type of meaning and semantic prosody. In view of the data examined, *deadly* and *mortal* were seen to have been reinterpreted over time

as degree modifiers, thus foregrounding degree. In this function, they modify unbounded heads and are scalar rather than totality modifiers, specifically boosters.

The data consulted revealed that the adverbial forms are the most grammaticalized, showing more intensifying uses than their adjectival counterparts. This is most conspicuous in the case of the adverb *mortal*, which was associated with an intensifying function in the most recent source materials, despite its low frequency in the data. Furthermore, in the case of this adverb, such an intensifying function correlated with the modification of positive heads. The more grammaticalized status of the two adverbs examined here may generally relate to the more prototypical function of adverbs as degree markers.

Although the data provided by some of the historical sources selected here, especially ARCHER, yielded very few intensifying tokens, they have led to important insights into the grammaticalization of *deadly* and *mortal* as intensifiers. Their intensifying function, though limited, suggests that they are pretty much *fashion victims*: they were most popular in LModE and then went out of use again, replaced by other more powerful intensifiers. In this regard, it may be interesting to explore the potential intensifying uses of *dead* and *mortally*, which were not considered here, to determine their frequency and potential uses as intensifiers, and whether they encroach on any of the functions of *deadly* and *mortal*. Likewise, the use of these death-related forms in other varieties of English may be worth examining. These issues, however, must remain for future research.

References

Corpora and software

- ARCHER = *A Representative Corpus of Historical English Registers* 3.2. 1990–1993/2002/2007/2010/2013. Originally compiled under the supervision of Douglas Biber & Edward Finegan at Northern Arizona University and University of Southern California. *Modified and expanded by subsequent members of a consortium of universities. Current member universities: Northern Arizona, Southern California, Freiburg, Heidelberg, Helsinki, Uppsala, Michigan, Manchester, Lancaster, Bamberg, Zurich, Trier, Salford, and Santiago de Compostela.* <www.alc.manchester.ac.uk/subjects/lel/research/projects/archer/>
- ECF = *Eighteenth Century Fiction*. Chadwyck Healey. 1996–2018. <http://collections.chadwyck.co.uk/home/home_c18f.jsp>
- EEBOCorp 1.0 = *Early English Books Online Corpus* 1.0, compiled by Peter Petré. 2013. <<https://lirias.kuleuven.be/handle/123456789/416330/>>
- HC = *The Helsinki Corpus of English Texts*. 1991. Department of Modern Languages, University of Helsinki. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English).

- MED = *Middle English Dictionary*. Sherman M. Kuhn, John Reidy & Robert E. Lewis (eds). 1952–2001. Ann Arbor, MI: University of Michigan Press. <<https://quod.lib.umich.edu/m/med/>>
- NCF = *Nineteenth Century Fiction*. Chadwyck Healey. 2000–2018. <http://collections.chadwyck.co.uk/home/home_c19f.jsp>
- OED = *Oxford English Dictionary* Online. Oxford: OUP. <<http://www.oed.com/>>
- WordSmith 6.0. Scott, Mike. 2012. *WordSmith Tools version 6*. Stroud: Lexical Analysis Software. <<https://lexically.net/wordsmith/>>

Other references

- Adamson, Sylvia. 2000. A lovely little example: Word order options and category shift in the premodifying string. In *Pathways of Change: Grammaticalization in English* [Studies in Language Companion Series 53], Olga Fischer, Anette Rosenbach & Dieter Stein (eds), 39–66. Amsterdam: John Benjamins.
- Aijmer, Karin. 2018. *That's well bad*: Some new intensifiers in spoken British English. In *Corpus Approaches to Contemporary British English*, Vaclav Brezina, Robbie Love & Karin Aijmer (eds), 60–95. New York NY: Routledge.
- Athanasiadou, Angeliki. 2007. On the subjectivity of intensifiers. *Language Sciences* 29(4): 554–565. <https://doi.org/10.1016/j.langsci.2007.01.009>
- Bolinger, Dwight. 1972. *Degree Words*. The Hague: Mouton.
- Breban, Tine & Davidse, Kristin. 2016. The history of *very*: The directionality of functional shift and (inter)subjectification. *English Language and Linguistics* 20(2): 221–249. <https://doi.org/10.1017/S1360674315000428>
- Cacchiani, Silvia. 2017. Cognitive motivation in English complex intensifying adjectives. *Lexis* 10: 1–21. <https://doi.org/10.4000/lexis.1079>
- Claridge, Claudia & Kytö, Merja. 2014. *I had lost sight of them for a bit, but I went on pretty fast*: Two degree modifiers in the Old Bailey Corpus. In *Diachronic Corpus Pragmatics* [Pragmatics & Beyond New Series 243], Irma Taavitsainen, Andreas H. Jucker & Jukka Tuominen (eds), 29–52. Amsterdam: John Benjamins.
- D'Arcy, Alexandra. 2015. Stability, stasis and change: The longue durée of intensification. *Diachronica* 32(4): 449–493. <https://doi.org/10.1075/dia.32.4.01dar>
- Fuchs, Robert. 2017. Do women (still) use more intensifiers than men? Recent change in the sociolinguistics of intensifiers in British English. *International Journal of Corpus Linguistics* 22(3): 345–374. <https://doi.org/10.1075/ijcl.22.3.03fuc>
- Heine, Bernd. 2003. Grammaticalization. In *The Handbook of Historical Linguistics*, Brian D. Joseph & Richard D. Janda (eds), 575–601. Oxford: Blackwell.
- Himmelman, Nikolaus P. 2004. Lexicalization and grammaticalization: Opposite or orthogonal? In *What Makes Grammaticalization: A Look from its Fringes and its Components*, Walter Bisang, Nikolaus P. Himmelman & Björn Wiemer (eds), 21–42. Berlin: Mouton de Gruyter.
- Hopper, Paul J. 1991. On some principles of grammaticization. In *Approaches to Grammaticalization, Vol. I: Theoretical and Methodological Issues* [Typological Studies in Language 19], Elizabeth Closs Traugott & Bernd Heine (eds), 17–35. Amsterdam: John Benjamins.
- Hopper, Paul J. & Traugott, Elizabeth Closs. 2003. *Grammaticalization*, 2nd edn. Cambridge: CUP.
- Lehmann, Christian. 1985. Grammaticalization: Synchronic variation and diachronic change. *Lingua e Stile* 20(3): 303–318.

- López-Couso, María José. 2010. Subjectification and intersubjectification. In *Historical Pragmatics*, Andreas H. Jucker & Irma Taavitsainen (eds), 127–163. Berlin: Mouton de Gruyter.
- Lorenz, Gunter. 2002. *Really worthwhile or not really significant?* A corpus-based approach to the delexicalization and grammaticalization of intensifiers in Modern English. In *New Reflections on Grammaticalization* [Typological Studies in Language 49], Ilse Wischer & Gabriele Diewald (eds), 143–161. Amsterdam: John Benjamins.
- Louw, Bill. 2000. Contextual prosodic theory: Bringing semantic prosodies to life. In *Words in Context: A Tribute to John Sinclair on his Retirement*, Chris Heffer & Helen Saunston (eds), 48–94. Birmingham: University of Birmingham.
- Macaulay, Ronald. 2006. Pure grammaticalization: The development of a teenage intensifier. *Language Variation and Change* 18(3): 267–283. <https://doi.org/10.1017/S0954394506060133>
- Méndez-Naya, Belén. 2006. Adjunct, modifier, discourse marker: On the various functions of *right* in the history of English. *Folia Linguistica Historica* 27(1–2): 141–169. <https://doi.org/10.1515/FLIH.2007.141>
- Méndez-Naya, Belén & Pahta, Päivi. 2010. Intensifiers in competition: The picture from early English medical writing. In *Early Modern English Medical Texts: Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 191–214. Amsterdam: John Benjamins.
- Nevalainen, Terttu. 2008. Social variation in intensifier use: Constraint on *-ly* adverbialization in the past? *English Language and Linguistics* 12(2): 289–315. <https://doi.org/10.1017/S1360674308002633>
- Nevalainen, Terttu & Rissanen, Matti. 2002. *Fairly pretty* or *pretty fair*? On the development and grammaticalization of English downtoners. *Language Sciences* 24(3–4): 359–380. [https://doi.org/10.1016/S0388-0001\(01\)00038-9](https://doi.org/10.1016/S0388-0001(01)00038-9)
- Palacios Martínez, Ignacio & Núñez-Pertejo, Paloma. 2012. *He's absolutely massive. It's a super day. Madonna, she's a wicked singer.* Youth language and intensification: A corpus-based study. *Text and Talk* 32(6): 773–796. <https://doi.org/10.1515/text-2012-0036>
- Paradis, Carita. 2008. Configurations, construals and change: Expressions of DEGREE. *English Language and Linguistics* 12(2): 317–343. <https://doi.org/10.1017/S1360674308002645>
- Partington, Alan. 1993. Corpus evidence of language change: The case of the intensifier. In *Text and Technology: In Honour of John Sinclair*, Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds), 177–192. Amsterdam: John Benjamins.
- Partington, Alan. 2015. Evaluative prosody. In *Corpus Pragmatics: A Handbook*, Karin Aijmer & Christoph Rühlemann (eds), 279–303. Cambridge: CUP.
- Peters, Hans. 1994. Degree adverbs in Early Modern English. In *Early Modern English*, Dieter Kastovsky (ed), 269–288. Berlin: Mouton de Gruyter.
- Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.
- Smith, K. Aaron & Nordquist, Dawn. 2012. A critical and historical investigation into semantic prosody. *Journal of Historical Pragmatics* 13(2): 291–312. <https://doi.org/10.1075/jhp.13.2.05smi>
- Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2(1): 23–55. <https://doi.org/10.1075/fol.2.1.03stu>
- Tagliamonte, Sali A. 2008. *So different* and *pretty cool*! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics* 12(2): 361–394. <https://doi.org/10.1017/S1360674308002669>
- Traugott, Elizabeth Closs. 1989. On the rise of epistemic meanings in English. An example of subjectification in semantic change. *Language* 65(1): 31–55. <https://doi.org/10.2307/414841>

- Traugott, Elizabeth Closs. 1995. Subjectification in grammaticalization. In *Subjectivity and Subjectivisation: Linguistic Perspectives*, Dieter Stein & Susan Wright (eds), 37–54. Cambridge: CUP.
- Traugott, Elizabeth Closs. 2003. From subjectification to intersubjectification. In *Motives for Language Change*, Raymond Hickey (ed.), 124–139. Cambridge: CUP.
- Traugott, Elizabeth Closs. 2007. (Inter)subjectification and unidirectionality. *Journal of Historical Pragmatics* 8(2): 295–309. <https://doi.org/10.1075/jhp.8.2.07clo>
- Traugott, Elizabeth Closs. 2010. (Inter)subjectivity and (inter)subjectification: A reassessment. In *Subjectification, Intersubjectification and Grammaticalization*, Kristin Davidse, Lieven Vandelanotte & Hubert Cuyckens (eds): 29–71. Berlin: Mouton de Gruyter.

So-called *-ingly* adverbs in Late Middle and Early Modern English

Yoko Iyeiri

Kyoto University

This chapter traces the increase of *-ingly* adverbs from Late Middle to Early Modern English. Not only the frequency and versatility but also the functions of *-ingly* adverbs expanded in this period: from the modification of adjectives, adverbs, and verbs (found both in the Middle and Early Modern English data) to the modification of the whole sentence (found only in the Early Modern English data). Sentential adverbs of the commenting type (disjuncts) are still absent throughout the entire period of this study. Finally, this chapter discusses some *-ingly* adverbs of the so-called “Harry Potter” type, suggesting that their development is earlier than assumed in previous studies.

1. Introduction

The present chapter explores the historical development of *-ingly* adverbs, where *-ly* has been attached to *-ing* forms of various kinds, as illustrated in Example (1).¹

- (1) ... whereunto I most *lovingly* invite thee
(EEBO: Christopher Farewell, 1633, *An East-India Colation*)

Lovingly in (1) illustrates the case in point. Adverbs of this type have not been investigated much to date, but thanks to some studies such as Kjellmer (1984), Killie (1998, 2000), and Broccias (2012a, 2012b), their overall development in the history of English has been brought to light. Whereas *-ingly* adverbs are very productive in Present-day English (cf. Killie 1998; Swan 1997, 2003), their establishment is considered to be fairly late. According to Killie (1998: 121; 2000: 44), *-ingly* adverbs were available in Old English, but their real expansion started only from around the 14th century. She gives a list of the *-ingly* adverbs she has found, which includes

1. Unless otherwise stated, all examples in this study are cited from *Early English Books Online* (EEBO), <<https://eebo.chadwyck.com/home>> (1 November 2018).

only *willingly*, *cunningly*, *hyingly* ('quickly'), and *overhyingly* before the 14th century (Killie 1998: 120–121). She also surveys the Middle and Early Modern English periods of the *Helsinki Corpus*, finding only 16 examples for the whole Middle English period, of which 10 are in the period 1350–1500, and 80 examples in the Early Modern English period (Killie 1998: 123). She concludes: "it [-ingly adverb] was not in common use until the Early Modern English period" (1998: 127). She even states: "From the 14th century on there has been a steady growth of new adverbs of this kind, but the category does not become truly productive until the 19th century" (2000: 44).

In recent years, adverbs of this particular form have been studied by Broccias (2012a, 2012b). He admits that Early Modern English shows an increased use of -ingly adverbs, as Killie (1998, 2000) remarks, but argues that those that expand at this early stage tend to be fixed ones like *willingly* and *accordingly*. He then suggests that the real expansion of -ingly adverbs took place in the Late Modern English period and that this is linked to the emergence of what he calls the "Harry Potter adverbs" – thus named as they occur commonly in Harry Potter novels. The following (cited from Broccias 2012a: 148; 2012b: 744) are illustrative examples of the -ingly adverbs of this kind:

- (2) "I'm not sure this is going to work, you know", said Hermione *warningly*.
(Rowling 2000: 228)
- (3) He seemed to be making some inquiry of the stranger, who shook his head *unsmilingly* and replied in an undertone.
(Rowling 2000: 164)

According to Broccias (2012a: 151), Harry Potter adverbs are "adverbs whose bases are clearly verbal" and typically occur with "verbs of saying, motion and watching". They consist of two types: (2) is an example of the "subjective" type, in which *warningly* gives the "subjective evaluation" of the event described by the main verb *said*, while (3) is an example of the "objective" type, in which *unsmilingly* gives the "objective description" of the event likewise described by the main verb (see Broccias 2012a: 148–151). He maintains that -ingly adverbs in earlier periods are not of the Harry Potter type, while those encountered from Late Modern English onwards are often of the Harry Potter type. The rise of Harry Potter adverbs is, according to him, the key to the increase of -ingly adverbs in the history of English.

Apart from these studies, not much research has been conducted into the historical development of -ingly adverbs. Killie's research into the historical development of adverbs is extensive, but does not pay much attention to -ingly adverbs. Broccias's studies focus on -ingly adverbs, but they are more concerned with the later period of Modern English, when the rise of Harry Potter adverbs is observed,

while his discussion of earlier days is based upon the *Helsinki Corpus*, which is rather concise by today's standard. Thanks to the development of technology, a much larger number of texts are analyzable today, which is more appropriate for research like the present one. I will focus upon the Late Middle and Early Modern English periods, which have been studied to a lesser extent. In this chapter, I will deal with all types of *-ingly* adverbs, so long as they end with *-ingly*. This is because the derivation is not always easy or even possible to judge. *Willingly* is an illustrative case in point. Broccias (2012a: 148) regards it as a derivative from the adjective *willing*, while Killie (1998: 120) considers the same item to be verbal in derivation.

The rest of this chapter is organized as follows: Section 2 describes the data explored in this study, Section 3 makes a survey of the overall increase of *-ingly* adverbs in Late Middle and Early Modern English, Section 4 discusses the types of *-ingly* adverbs attested in this study, Section 5 deals with various functions of *-ingly* adverbs, and the final section concludes the entire discussion of this chapter.

2. Data

With the aim of highlighting the rise of *-ingly* adverbs from Late Middle to Early Modern English, I have explored three corpora, which I have created by assorting some of the transcribed texts from *Early English Books Online* (EEBO): (a) *Selected Middle English Texts in Print* (METiP) (a collection of 15th-century printed texts, mainly but not exclusively prose); (b) *Selected Authors' Writings in the Early Modern English Period* (SAEMEP), from which I have extracted the works of Sir Thomas More (1478–1535), Francis Bacon (1561–1626), Richard Baxter (1615–1691), and John Locke (1632–1704); and (c) *Early Modern English Prose Selections, A-Texts* (EMEPA-A) (version 2) (1501–1700). The analysis in this paper is essentially based on (a) and (b), while (c) is used as a reference corpus. In the remainder of this section, I will give some details of the corpora.

METiP (a) is a collection of printed texts in the 15th century. It consists of five groups of Middle English texts written by some major authors: Geoffrey Chaucer (approximately 321,100 words), John Trevisa (approximately 364,900 words), Thomas Malory (approximately 350,100 words), William Caxton (approximately 353,100 words), and some religious texts (approximately 358,000 words). Table 1 displays the basic structure of this collection.

Table 1. Selected Middle English Texts in Print (METiP)*

Author/ Topic domain	Texts	Approximate number of words
Geoffrey Chaucer	Six texts (prose and verse) printed 1477–1500	321,100
John Trevisa	<i>Prolicionycion</i> [sic], printed in 1482	364,900
Thomas Malory	<i>Le Morte Darthur</i> , printed in 1485	350,100
William Caxton	Five texts printed 1481–1485	353,100
Religious Texts	Texts by Walter Hilton, Nicholas Love, Richard Fitzjames, Simon Winter, and Raymond of Capua, printed 1494–1500	358,000
Total		1,747,200

* Iyeiri (2016) provides a full list of the texts in METiP.

The texts included in METiP go back to the last quarter of the 15th century, although some of them date further back to the 14th century in terms of their original composition. As Table 1 shows, METiP is not a balanced corpus like the *Helsinki Corpus*. Hence, it is not exactly a powerful dataset for genre analyses, although the group of Religious Texts may reveal some characteristic features of this genre. On the other hand, METiP is fairly substantial, amounting to about 1.7 million words. It will possibly demonstrate more clearly than in previous studies the overall state of the development of *-ingly* adverbs in Late Middle English.

Secondly, SAEMEP (b) is a collection of texts written by four selected individuals in the Early Modern English period. As Table 2 shows, (b) is a collection of about two million words, with approximately 500,000 words each for More, Bacon, Baxter, and Locke. This dataset is, like METiP, not exactly suitable for strict genre studies, but will demonstrate the process of the development of *-ingly* adverbs in Early Modern English, which needs further research as stated in Section 1. The result obtained from SAEMEP will be compared with *-ingly* adverbs in EMEPS-A (c),

Table 2. The texts written by More, Bacon, Baxter, and Locke in SAEMEP*

Authors	Texts	Approximate number of words
Sir Thomas More (1478–1535)	Seven texts by Sir Thomas More	462,600
Francis Bacon (1561–1626)	Ten texts by Francis Bacon	521,000
Richard Baxter (1615–1691)	Eleven texts by Richard Baxter	504,200
John Locke (1632–1704)	Seven texts by John Locke	517,200
Total		2,005,000

* SAEMEP, which is still under construction, comprises groups of texts written by single authors, including Thomas More, Francis Bacon, Richard Baxter, John Locke, and others. The word counts in Table 2 are approximate, but based upon the texts with inappropriate portions such as advertisements excluded manually. Further details of this collection are given in the Appendix.

Table 3. Early Modern English Prose Selections, A-Texts, version 2 (EMEPA-A)*

	Files	Texts	Approximate number of words
A-texts	A16th-1 (1500–1550)	43 texts by various authors	1,013,600
	A16th-2 (1551–1600)	53 texts by various authors	1,002,400
	A17th-1 (1601–1650)	95 texts by various authors	1,002,200
	A17th-2 (1651–1700)	75 texts by various authors	1,001,200
Total			4,019,400

* EMEPS-A has now been updated (version 2) by eliminating the text by Cristiern Pedersen (1533) and adding instead Miles Coverdale's *A Confutacion of that Treatise* (1541). Hence the slight difference in the number of words between EMEPS (version 1) and EMEPS (version 2).

wherever appropriate, which is slightly different in respect of the corpus design from METiP and SAEMEP. EMEPS-A is a much larger corpus and a more general one, including a number of texts written by numerous different authors. Thus, (a) and (b) form the central data in this study, while (c) functions as a reference corpus in the discussion below.

Since I have already given a full account of EMEPS in Iyeiri (2011), which is easily accessible online, I will only give a brief description of it here. EMEPS consists of prose texts written by a number of authors in the period 1501–1700 and is grouped into A-texts and B-texts, both of which are further grouped into four sub-periods, 1501–1550, 1551–1600, 1601–1650, and 1651–1700 respectively.² In this study, I will use the A-texts only, whose details are shown in Table 3. EMEPS-A consists of approximately four million words. Although both (b) and (c) use EEBO as their source, they do not have any duplication of texts. In other words, (b) and (c) are mutually exclusive.

3. The increase of *-ingly* adverbs in Late Middle and Early Modern English

As mentioned in Section 1, the use of *-ingly* adverbs is, thanks to previous studies, known to have expanded at least from the 14th century onwards. They are indeed attested in the three corpora I have explored. Figure 1 shows how frequently they are attested (per million words) in: (a) Chaucer, Trevisa, Malory, Caxton, and Religious Works in METiP; and (b) More, Bacon, Baxter, and Locke in SAEMEP.

2. The purpose of dividing EMEPS into EMEPS-A and EMEPS-B was simply to make the corpus more manageable. Some research questions require eight million words (EMEPA-A and EMEPS-B), while for others four million words (either EMEPS-A or EMEPS B) are quite sufficient. The essential corpus design is the same between A and B. For further details, see Iyeiri (2011).

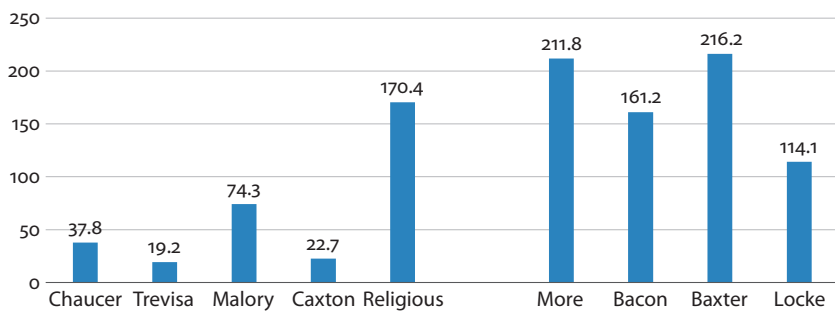


Figure 1. Frequencies of *-ingly* adverbs in METiP and four major authors in SAEMEP (per million words)

Some illustrative examples are:

- (4) Thyse ben the newe shepherdes that leten her sheep *wytyngly* go renne to the wolf that is in the breris and do no force of her owen gouernance (Chaucer, 1477, *Canterbury*)
- (5) Than a crowe that she hadde *lykyngly* fedde and norysshed creaked ludder than he was wonted (Trevisa, 1482, *Polychronicon*)
- (6) For this is the vsage of the wyse woman / whiche tendeth to lyue peasybly and *louyngly* with her husbond and lord (Caxton, 1484, *La Tour*)
- (7) From which they differ *exceedingly*, as much as Wine from Water; And as Wood from Earth. (Bacon, 1627, *Sylua syluarum*)
- (8) That they who can see no Indications of a changeable Nature, must needs be *willingly* ignorant (Locke, 1694, *Reason and Religion*)

As Figure 1 demonstrates, *-ingly* adverbs are not altogether common in the Middle and Early Modern English periods, but the sudden expansion of their use around the end of the Middle English period or the beginning of the Modern English period is noteworthy. Although the rise is not necessarily regular, the transitional period from Middle to Early Modern English is clearly a key period in terms of the development of *-ingly* adverbs. METiP provides 114 examples in about 1,747,200 words, while the four major authors in SAEMEP yield 350 examples in about 2,005,000 words.

Although the datasets of the present study are not necessarily suitable for genre analyses, the graph suggests that there is probably a slight tendency for religious writings to show *-ingly* adverbs more extensively than texts of other types. The statistics above display the relatively frequent occurrences of *-ingly* adverbs in Religious Texts in METiP, and also in Thomas More and Richard Baxter in SAEMEP, the former being a pious Catholic while the latter a Puritan church leader. This may

be partly due to the fairly common use of *everlastingly* in religious texts, which occurs only in the Religious Texts in METiP and More and Baxter in SAEMEP. Further details of the frequencies of different types of *-ingly* adverbs are given in the next section.

The overall tendency, as shown in Figure 1, is more or less in keeping with EMEPS-A, which is a reference corpus. The frequencies of *-ingly* adverbs per million words in EMEPS-A are: 158.8 in 1501–1550, 270.4 in 1551–1600, 239.5 in 1601–1650, and 178.8 in 1651–1700. The rise and fall of the frequencies is not necessarily regular, either, but they are within the range of 150 to 280.³ This situation is similar to that of SAEMEP. The use of *-ingly* adverbs is clearly more extensive in Early Modern English than in Middle English. All in all, it is safe to state that the period from Late Middle to Early Modern English is intriguing in terms of the expansion of *-ingly* adverbs, while most previous studies tend to focus on the period from Late Modern English to the present day.

4. Different types of *-ingly* adverbs

Whereas the overall rise of *-ingly* adverbs from Late Middle to Early Modern English is transparent as discussed above, different authors seem to present a significantly different state of affairs, not only in the frequencies but also in the diversity of the *-ingly* adverbs they employ. *Wittingly* (including its orthographic variants such as *wetyngly*) is the only *-ingly* adverb shared by all of the five subgroups of METiP (Chaucer, Trevisa, Malory, Caxton, and Religious Texts).⁴ Apart from it, different authors display different preferences in the types of *-ingly* adverbs.⁵ Malory, for example, is quite distinct from the others, in that he gives only two

3. Although the frequencies of *-ingly* adverbs in the Early Modern English period are all within the range of 150 to 280, there seems to be a notable rise from the first half to the second half of the 16th century. I have noted in one of my earlier publications that this is a crucial time period in the history of English, when a number of linguistic features experienced a sudden spike. See Iyeiri (2014: 38) for an outstanding increase of *always* (as against *alway*) in this period. Raumolin-Brunberg and Nevalainen (1994) also state that some major linguistic changes, including the peaking of *of*-genitive instead of *s*-genitive and the verbal ending *-s* instead of *-th*, took place around this time.

4. Unless otherwise stated, all orthographic variants are included under the quoted forms. This practice is followed throughout this study.

5. Although original texts may have been altered in the process of textual transmission and preparation for printing, I will refer, for practical reasons, to Chaucer, Trevisa, and Malory by their names in this study.

types: *passingly* (25 examples) and *wittingly* (one example).⁶ The present section pays attention to different types of *-ingly* adverbs.

On the whole, Middle English authors are inclined to show little diversity in their inventory, though the number of types of *-ingly* adverbs is much larger than suggested in previous studies. See Table 4 for some details of the *-ingly* adverbs in METiP.

Table 4. Occurrences of *-ingly* adverbs (types and tokens) in METiP*

Author/ Topic domain	Types (tokens)	Totals
Chaucer	<i>wetyngly</i> (3), <u><i>brennyngly</i></u> (2), <i>bityngly</i> (1), <i>cunnyngly</i> (1), <i>felyngly</i> (1), <i>sokyngly</i> (1), <u><i>vnfelyngly</i></u> (1), <i>vnwetyngly</i> (1), <i>willyngly</i> (1)	12
Trevisa	<i>lykyngly</i> (3), <u><i>feynyngly</i></u> (1), <u><i>flatelyngly</i></u> (1), <u><i>passyngly</i></u> (1), <i>wyttyngly</i> (1)	7
Malory	<u><i>passyngly</i></u> (25), <i>wetyngly</i> (1)	26
Caxton	<i>wytyngly/wetyngly</i> (2), <i>vnwytyngly/vnwetyngly</i> (2), <u><i>dyssimylyngly</i></u> (1), <i>louyngly</i> (1), <u><i>perceyuyngly</i></u> (1), <i>wyllyngly</i> (1)	8
Religious Texts	<i>euerlastyngly</i> (8), <i>lykyngly/lykyngly</i> (8), <i>lastyngly</i> (6), <i>passingly/passyngly</i> (5), <u><i>abidingly/abydyngly</i></u> (4), <u><i>brennyngly</i></u> (4), <i>louyngly</i> (4), <u><i>restyngly</i></u> (4), <i>wyttyngly</i> (4), <i>felyngly</i> (3), <i>folowyngly</i> (3), <u><i>plesyngly</i></u> (2), <u><i>smylyngly</i></u> (2), <u><i>cleuyngly</i></u> (1), <u><i>feynyngly</i></u> (1), <i>stiryngly</i> (1), <i>welyngly</i> (1)	61
Total		114

* For reference's sake, the items underlined in this list are those not attested in Killie's (1998: 120–121) list of *-ingly* adverbs.

Religious Texts provide a larger number of types than Chaucer, Trevisa, Malory, and Caxton, despite the more or less equal size of the five sub-collections. I have shown earlier that the common occurrence of *everlastingly* can be one of the reasons for the frequent attestation of *-ingly* adverbs in general in this genre. This may be true to some extent, but it is not only the overall frequency that is remarkable in this group. As Table 4 reveals, the number of different types of *-ingly* adverbs is also large in this subgroup.

One possible explanation for this diversity would be that the nature of this particular subgroup is slightly different from the other groups within METiP. Unlike Chaucer, Trevisa, Malory, and Caxton, it is a collection of texts written by several different authors. Table 4 shows that different authors tend to have different sets of

6. As *wittingly* (with the orthographic form *wetyngly*) is the only form, apart from *passingly*, used by Malory, I have double-checked it in the Winchester Malory by use of Ker's (1976) facsimile edition, where it is available, though in the spelling *wittyngly* (223v). Hence, it is not a printers' introduction.

-ingly adverbs. It is quite possible that the number of authors involved is relevant to the number of the types of *-ingly* adverbs. In other words, the more authors involved, the larger the number of types that is expected. One is reminded of the fact that *wittingly* is the only form shared by the five subgroups of METiP. It is among the *-ingly* adverbs first attested in the 14th century according to Killie (1998: 120), namely among the oldest. *Passingly*, which is characteristic of Malory and which is shared by Trevisa and Religious Texts as well, may also have been well-established in Late Middle English, considering the fact that even Malory's work, which is essentially devoid of *-ingly* adverbs, shows its attestation. It is not in the list of Killie, as the verb *pass* was outside the purview of her investigation. Apart from these, *everlastingly* (in Religious Texts only), *likingly* (in Chaucer and Religious Texts), and *lastingly* (in Religious Texts only), which are relatively common, may have been relatively well established in Late Middle English.⁷

In fact, the *Middle English Dictionary* (*MED*) gives separate entries for most of the *-ingly* adverbs listed in Table 4. The adverbs whose citations in the *MED* date back to the 14th century are: *brenningly* (1340), *clevingly* (1340), *lastingly* (c1340), *passingly* (1340), *wittingly* (1340), *everlastingly* (c1350), *cunningly* (1357), *feelingly* (a1382), *followingly* (a1382), *stirlingly* (a1382), *unwittingly* (a1382), *unfeelingly* (c1385), *feiningly* (a1387), *flatteringly* (a1387), *likingly* (a1387), *willingly* (c1395), *restingly* (a1396), *pleasingly* (a1398), *smilingly* (a1400), and *bitingly* (?a1400). Those whose quotations date back to the 15th century are: *abidingly* (a1410), *lovingly* (a1420), and *seemingly* (?a1425).⁸ Those commonly attested in METiP, namely *wittingly*, *passingly*, *everlastingly*, *likingly*, and *lastingly*, all present examples dating back to the 14th century in the *MED*. Overall, the list from the *MED* also proves that *-ingly* adverbs were not as uncommon in Middle English as assumed in previous studies.

When attention is directed to the Early Modern English period, a clearer contrast exists between those adverbs with high frequencies and those with low frequencies. Some *-ingly* adverbs are fully established, presenting a notable number of examples, while others are more occasional (or even ad hoc). See Figures 2 to 5.⁹

7. There is no overlap between *everlastingly* and *lastingly* in the data of this study. Examples of *lastingly* include: *Thou shalte haue the meanyng of thise wordes in thyn entent & in habytt of thy soule lastyngly* (Hilton, 1494, *Scala perfeccionis*).

8. Here I simply indicate the earliest dates provided by the *MED*, overviewing the rise of *-ingly* adverbs in the Middle English period. Some are (estimated) dates of the original composition, while others are dates of the manuscript.

9. I have decided that for the data of Early Modern English, the graphical presentation will reveal more clearly the contrast between frequent *-ingly* adverbs and infrequent ones.

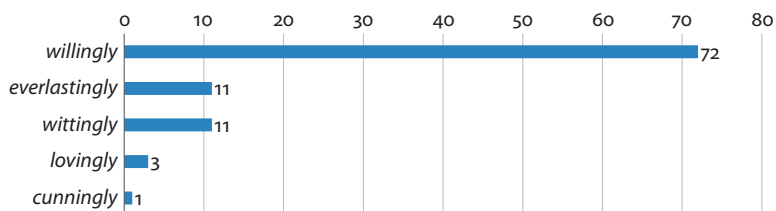


Figure 2. *-ingly* adverbs in Sir Thomas More (raw frequencies)

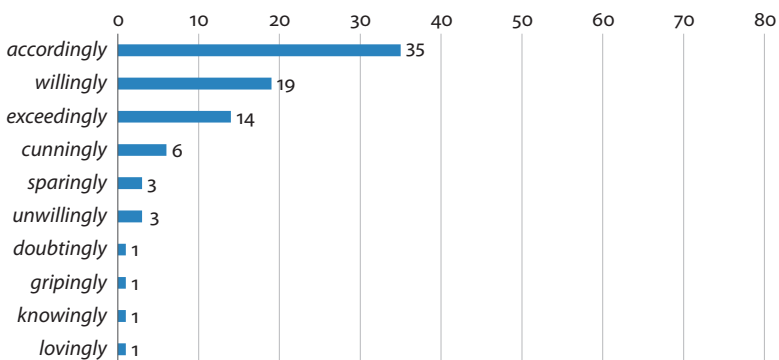


Figure 3. *-ingly* adverbs in Francis Bacon (raw frequencies)

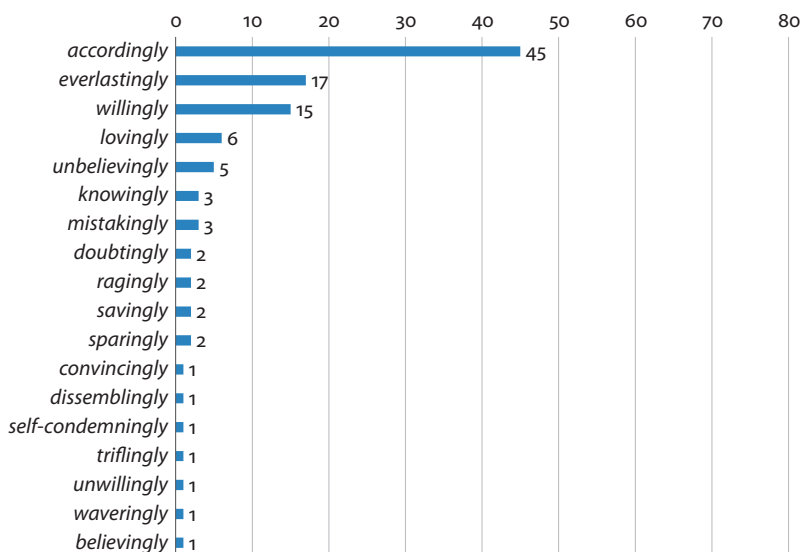


Figure 4. *-ingly* adverbs in Richard Baxter (raw frequencies)

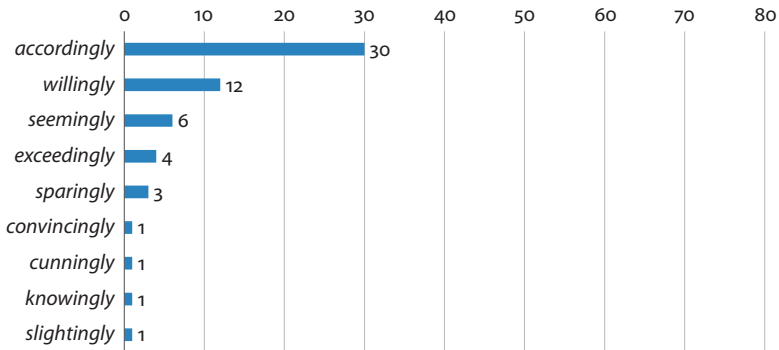


Figure 5. *-ingly* adverbs in John Locke (raw frequencies)

Thomas More's English illustrates a transitional stage from Middle to Early Modern English (see Figure 2): his use of *-ingly* adverbs is largely confined to well-established types such as *willingly*, *everlastingly*, *wittingly*, and *lovingly*, and hardly extends to occasional types with low frequencies. *Cunningly*, occurring only once, is probably the sole example of the latter category in his works.

By contrast, the language of later authors (Figures 3–5) displays further diversity of *-ingly* adverbs and, consequently, the contrast between fully-established adverbs and rare ones is more striking, although ultimately, it is impossible to draw a clear boundary between the two kinds of adverbs. Figures 3–5 suggest that *accordingly*, *willingly*, *everlastingly*, *exceedingly*, *seemingly*, *lovingly*, and perhaps *cunningly* have been well established with high frequencies, among which *accordingly* is quite noticeable in that its use is predominantly from Bacon onwards. It is hence an adverb whose use expanded fairly suddenly in the Early Modern English period.

The overall picture attained in this research is largely in accordance with Broccias's (2012a: 151–152) study of the *Helsinki Corpus*, which yields *willingly* (24 examples), *accordingly* (21 examples), *exceedingly* (11 examples),¹⁰ *cunningly* (4 examples), *sparingly* (3 examples), *wittingly* (3 examples), *lovingly* (2 examples), *seemingly* (2 examples), *everlastingly* (2 examples), and others (1 example each). Of the adverbs listed here, *accordingly*, *willingly*, *exceedingly*, and *cunningly* are relatively common in the present study in general, while *everlastingly* and *lovingly* are relatively common in Religious Texts: they are distinctively frequent in Baxter,

10. Broccias's table (2012a: 152) deals with different orthographic forms as separate forms, while I have combined *exceedingly*, *exceedingly*, and *excedingly* as one form. The same applies to *euerlastingly* and *everlastingly*.

and they are two of the five types of *-ingly* adverbs in More's writings.¹¹ On the other hand, *seemingly* appears to be a relatively late type, as it is observed only in Locke, who belongs to the youngest generation of the four. The *OED* provides two examples earlier than the 17th century under the entry of *seemingly*, but all the other examples are later.¹²

Overall, the above description supports Killie (1998: 123), who regards "the Middle English era as the period when the pattern became established, and the Early Modern era as the period when it began to be used". It is quite likely that *-ingly* became a clearly productive category for many authors by the time of the Early Modern English period, considering the fact that they tend to present fairly numerous rather occasional examples as well as well-established ones.

For the sake of reference, I will present the data of the first half of the 17th century in EMEPS-A (Figure 6). This is to confirm the tendency observed in SAEMEP. The versatility of *-ingly* adverbs is even clearer than in SAEMEP, which is likely due to the inclusion of multiple authors in EMEPS-A. In this corpus, too, the contrast between well-established *-ingly* adverbs, that is *willingly*, *accordingly*, *exceedingly*, *cunningly*, etc. in this case, and those used in an occasional manner is clearly visible. This reconfirms that the Early Modern English period sees the full establishment of *-ingly* adverbs of traditional types and simultaneously the expansion of copious other forms, whose use depends on the creativity of individual authors.

An obvious question is whether the prominent rise of *-ingly* adverbs in this period is related to the rise of *-ly* adverbs in general. Nevalainen (1994: 244) discusses the rise of *-ly* adverbs in the history of English, and shows that their expansion is evidenced from Late Middle English onwards. Assuming that the beginning of the Early Modern English period is the key, I have investigated the five relevant forms

11. As the main body of the discussion reveals, *accordingly* is a very frequent adverb in Early Modern English. It is worthy of note, however, that Broccias (2012a: 155) refers to the same adverb and states that it "decreases substantially as we move from the 18th to the 19th century". The inventory of popular forms fluctuates depending on the date. To give additional examples, *passingly* and *wittingly* are among the popular forms in Late Middle English, whereas their entries in the *OED* do not present any examples in the 20th century. This is another case to illustrate the point.

12. I have explored the entire dataset of the *OED* on CD-ROM (version 4), obtaining the same result. This, however, does not deny the existence of *seemingly* in earlier periods. As mentioned above, the *MED* has an entry of this word, giving some examples. Earlier examples tend to bear a slightly different meaning ('in a proper manner' and 'befittingly') from the common use of the same word in Modern English (see *MED*, s.v. *semingli* and *OED*, s.v. *seemingly*).

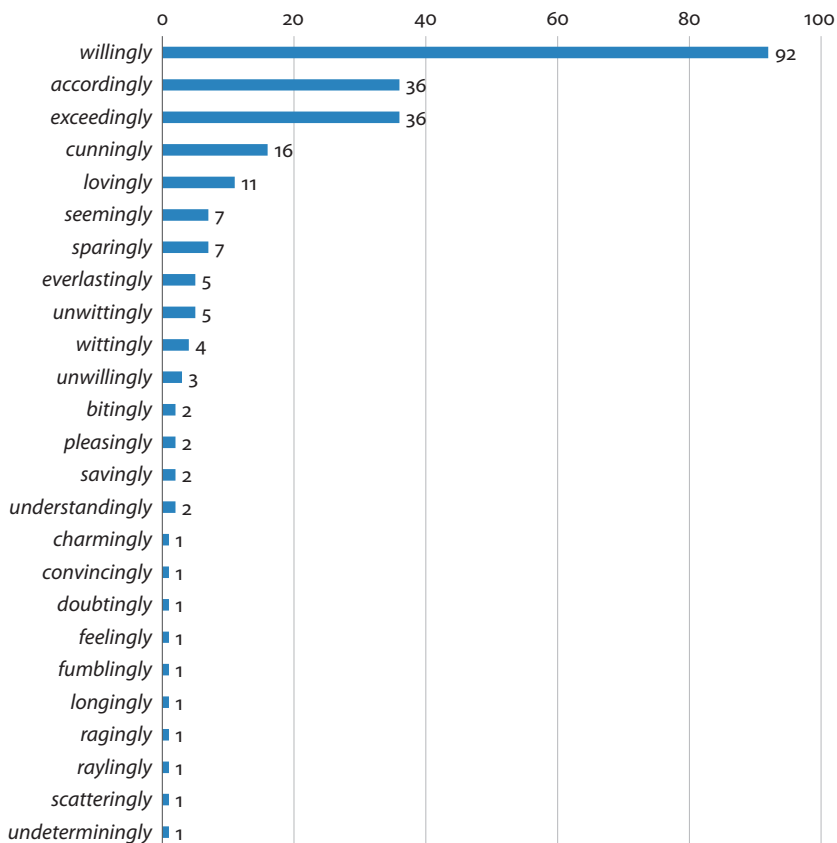


Figure 6. *-ingly* adverbs in 1601–1650 in EMEPS-A (raw frequencies)

without *-ly* in More's texts, that is *willing*, *everlasting*, *witting*, *loving*, and *cunning*, and found that all are adjectival. This is certainly an area that needs further exploration, but the result of this preliminary study shows that there does not seem to be any clear competition between *willing* and *willingly*, *everlasting* and *everlastingly*, etc. in the adverbial domain.¹³

13. This does not deny the possible relationship between the rise of *-ly* adverbs in general in the Early Modern English period and the rise of *-ingly* adverbs. It is quite possible that the environment where *-ly* was commonly used facilitated the increased use of *-ingly* adverbs, even though there was no competition between *willing* and *willingly*, etc.

5. Functions of *-ingly* adverbs

5.1 Modification of adjectives, verbs, and clauses

This section delves into different functions of *-ingly* adverbs in the Late Middle and Early Modern English periods. Adverbs are multi-functional, in that they can modify various different items in sentences: adjectives and other adverbs, verbs and verb phrases, and the entire sentence. I will deal with these three types of adverbs in the following discussion. Although the distinction between the modification of the verb or verb phrase and the entire sentence may not necessarily be clear when the verb is finite, there are two types of adverbs that are clearly sentential. They are what Quirk et al. (1985: 440–441) call “disjuncts” and “conjuncts”, corresponding respectively to “stance adverbials” and “connecting adverbials” in Biber et al. (1999: 549). The function of the former is to comment on the content of the whole sentence, while that of the latter is to connect sentences (or other elements). In the phrasing of Quirk et al. (1985: 440), disjuncts are adverbs “with speaker’s authority for, or comment on, the accompanying clause”, whereas conjuncts “express the speaker’s assessment of the relation between two linguistic units”. Hence, it is necessary to single out disjuncts and conjuncts as sentential adverbs, although in practice they are still very rare in the data of this study. I will, therefore, employ the terms disjuncts and conjuncts only when necessary, while the following discussion is mostly concerned with the modification of adjectives and adverbs on the one hand and modification of verbs and verb phrases on the other.

In the data of this study, most *-ingly* adverbs modify verbs and verb phrases, as illustrated in Examples (9) and (10).

- (9) Also whan one synneth *wytyngly* and in earnest. he synneth more than whan one synneth ygnorantlye. (Caxton, 1485, *Le liure royal*)
- (10) Those that write so hotly and *ragingly* to provoke Superiors to ruine us. (Baxter, 1689, *Cain and Abel*)

In some cases, the verb modified by *-ingly* can be non-finite, as in (11) and (12). In (11), *willingly* modifies the verb *say*, which is an infinitive, while in (12) *sparingly* modifies *living*, which is a present participle.

- (11) by cause that none shold be greued to say it *wyllyngly* and ofte. (Caxton, 1485, *Le liure royal*)
- (12) Monkes and Hermites living *sparingly* on drye foode, commonly attayned to a great age. (Bacon, 1638, *The Historie of Life and Death*)

Although *-ingly* adverbs modifying verbs and verb phrases are the principal kind both in Middle and Modern English texts of the present study, there are a handful

of *-ingly* adverbs used as local modifiers of adjectives and adverbs in METiP and SAEMEP. In METiP, which provides a total of 25 different types of *-ingly* adverbs, *passingly* is the only one that clearly belongs to this category. Unlike other *-ingly* adverbs in Middle English, *passingly* typically modifies adjectives and adverbs, as in (13)–(15).

- (13) for the kyng was *passyngly* wroth for the hurte of sir Gryflet
(Malory, 1485, *Morte Darthur*)
- (14) For this is more worthy and more to commende than a man be *passyngly*
stronge. and a co~querour of Cytees and londes. as Salamo~ witnessyth.
(Nicholas Love (religious), 1494, *Incipit speculum vite cristi*)
- (15) for that prynce was *passyngly* wel byloued ... (Malory, 1485, *Morte Darthur*)

METiP provides only a few examples of *passingly* that modify verbs and verb phrases. Examples include:

- (16) Therfore ther shall no man meruayle though I loue them *passyngly* to whome
our lord hath bede me warne them of theyr synfull lyuyng
(Raymond of Capua (religious), 1500, *The Lyf of Saint Katherin of Senis*)

Interestingly enough, the use of *-ingly* adverbs modifying adjectives and adverbs is virtually confined to Malory: as many as 25 of the 31 examples of *passingly* in METiP are attested in this text. Furthermore, Malory is virtually devoid of other *-ingly* adverbs, the sole exception being *wetyngly* (see Section 3). Also, Malory's use of *passingly* is comparatively fixed as it modifies only *sore* (five examples), *wroth* (two examples), and *wel* (18 examples). All these features put Malory in a unique position, in that his *-ingly* adverbs are limited to fixed forms whereas for other Middle English authors, they are already a productive category and their typical function is to modify verbs and verb phrases. Disjuncts and conjuncts are not yet evidenced in the Middle English texts explored in the present study.

The strong tendency for *-ingly* to modify verbs and verb phrases continues in the Early Modern English texts of SAEMEP. As for adverbs modifying adjectives and adverbs, *passingly*, which is the sole item used in this function in METiP, is no longer present in SAEMEP.¹⁴ Instead, *exceedingly* and *seemingly*, as shown below, are found to modify local constituents. They are, however, evidenced only in Locke's writings, while the other authors of SAEMEP do not provide any examples of *-ingly* adverbs modifying adjectives or adverbs, as in (17) and (18).

14. This does not necessarily mean that *passingly* had disappeared by this time. See the *OED* (s.v. *passingly*), which cites some relevant examples, though not numerous, from Modern English texts.

- (17) ... to give Mr. Edwards (who is *exceedingly* quick-sighted and positive in such Matters) a handle
(Locke, 1697, *A Second Vindication of the Reasonableness of Christianity*, &c)
- (18) ... such are the *seemingly* positive terms of Old, Great, Imperfect, &c.
(Locke, 1690, *An Essay Concerning Humane Understanding*)

The use of *seemingly* in (18) is interesting, as it reveals the judgement of the author, though it is used to modify adjectives locally. This may be a relatively new feature in the history of English.¹⁵

Thus, *exceedingly* and *seemingly* are the only items showing local modification of adjectives and adverbs, and they are confined to Locke. This does not imply, however, that this is the only attested usage of these adverbs. *Exceedingly* at least is found in contexts where it modifies verbs and verb phrases in Bacon (see (19)–(20)).

- (19) And therefore light rubbings and exercise breathing the body, not procuring sweate, doe *exceedingly* lengthen life.
(Bacon, 1638, *The Historie of Life and Death*)
- (20) So having written to your Lordship, I desire *exceedingly*, to be preferred, in your good Opinion, and Love.
(Bacon, 1657, *Resuscitatio*)

Bacon gives as many as 14 examples of *exceedingly*, all of which modify verbs and verb phrases, although there are some ambiguous examples such as “... was exceedingly moued” (1604, *Sir Francis Bacon his Apologie*) and “...was exceedingly contented” (1618, *A Declaration of the Demeanor and Cariage of Sir Walter Raleigh*), which can be interpreted as an example of “-ingly plus an adjective”.

As hitherto discussed, -ingly adverbs are essentially modifiers of verbs and verb phrases in Early Modern English, as in Late Middle English. This conclusion drawn from the analysis of SAEMEP is largely confirmed by the data of EMEPS-A. As for -ingly adverbs modifying adjectives and adverbs, EMEPS-A includes one example each of *convincingly* and *mountingly*, in addition to *exceedingly* and *seemingly* (see (21)–(24)).

- (21) Theyr paynes shall be so *excedyngly* greuous
(John Fisher, 1532, *Here After Ensueth Two Fruytfull Sermons*)
- (22) For his *seemingly* rigid Zeal for Religion was tempered with such an affability of deportment (Lanelot Addison, 1679, *The Life and Death of Mahumed*)

15. One example of *seemingly* in Locke modifies a noun phrase, and this again displays the judgement of the author: “In his gracious Allowance, Mr. Bold is *seemingly* a Man of some relish of Religion and Piety” (Locke, 1697, *A Second Vindication of the Reasonableness of Christianity*, &c). This usage is attested in A17–2 (1651–1700) of EMEPS-A as well. See also footnote 12.

- (23) ... and if these things which are written, be not yet *convincingly* satisfactory to each ingenuous, and non-pre-engaged Reader
(John Ellyson, 1647, *Hereticks, Sectaries, and Schismatics*)
- (24) ... either *mountingly* eloquent, or profoundly learned
(John Eachard, 1672, *The Grounds & Occasions of the Contempt of the Clergy and Religion Enquired into in a Letter Written to R.L.*)

Considering the fact that there are more than 70 different -ingly adverbs in total in the four authors of SAEMEP and EMEPS-A together, modifiers of adjectives and adverbs (*exceedingly*, *seemingly*, *convincingly*, and *mountingly*) are a minority.

One notable feature of the Early Modern English period is the rise of *accordingly*, which can, but not exclusively, function as a connector, that is a conjunct. This is a new usage found in the Early Modern English data.¹⁶ Examples include (25).

- (25) And *accordingly* this Noble Proselyte gives this account of his Faith, in order to his being Baptiz'd, in order to his being admitted a Member of Christ's Church, I believe that Jesus is the Son of God (Locke, 1697, *A Second Vindication of the Reasonableness of Christianity, &c*)

When it appears at the beginning of the sentence, it is often preceded by the conjunction *and*. When it occurs in the medial position, it usually modifies the verb-predicate, but can, though not always, refer conceptually to the previous proposition, as in (26).¹⁷

- (26) Whereupon the King (finding himselfe too weake) sought vnto King HENRY, as also vnto the Pope, and the King of France, to compose those troubles, betweene him and his Subiects. The Kings *accordingly* interposed their Mediation in a round and Princely manner (Bacon, 1629, *The Historie of the Reigne of King Henry the Seuenth*)

By contrast, *accordingly* refers to what follows, when it is accompanied by the preposition *to* or the conjunction *as*. (27) and (28) are cited from EMEPS-A, where relevant examples are more numerous than in SAEMEP.¹⁸

16. The *MED* still defines *accordingly* as "(a) in agreement (with sb. or sth.); in conformity or accordance (with); accordingly; ?also, approvingly; (b) properly; adequately".

17. *Accordingly* is polysemous and does not always have to imply a logical relationship to the previous proposition. The *OED* (s.v. *accordingly*) gives, for example, the meanings 'harmoniously', 'suitably', etc., which can stand independently of the preceding proposition.

18. This is simply a matter of 'relative' frequency. Thus, examples with *to* or *as* are by no means abundant in EMEPS-A, either. There is a slight tendency for earlier texts to provide more examples. A16-1 (1501-1550) gives eight examples (mostly with *to*), while it gives a total of 36

- (27) Yf Chrisostome ment *accordingly* to the ordinaunce of holy chirche we will accepte him.

(William Thorpe, 1530, *The Examinacion of Master William Thorpe*)

- (28) As touching the quantitie and measure of such remedies as hee must vse, he may as occasion serueth, (*accordingly* either as the age of the partie, or as his custome and ordinary dealings and fashions do require) occupy and take

(Thomas Newton, 1586, *The Olde Mans Dietarie*)

The rise of *accordingly* in Early Modern English, particularly the rise of the newly-acquired conjunct use, is worthy of note. This development is largely in conformity with what Swan (1991) calls “adverbial shifts”, by which adverbials in general display an increasing tendency to modify sentences rather than words.¹⁹ The shifts do not have to take place individually with all adverbials, but the over-all proportion of adverbs modifying the entire sentence increases as time passes. *Accordingly* is certainly one adverb of this kind. Its expansion in the Early Modern English period makes a significant contribution to, and is simultaneously supported by, the general trend of this shift.

On the other hand, sentential *-ingly* adverbs of the present study are not yet disjuncts in the real sense. As mentioned above, *seemingly* conveying the author’s or speaker’s judgement in some 17th-century materials merits attention. They do not, however, function in the same way as adverbs like *surprisingly* in contemporary English, which can be more clearly sentential.

5.2 Harry Potter adverbs

This study, except in Section 1, has so far dealt with *-ingly* adverbs without referring to Harry Potter adverbs, which according to Broccias are attested “from the beginning of the 19th century” (2012b: 744–745). Concerning *-ingly* adverbs in Early Modern English, Broccias states:

... Killie (1998) shows that *-ingly* adverbs established themselves in the EModE period but, on closer inspection, this observation turns out not to apply to the Harry Potter type. The *-ingly* adverbs Killie refers to as emerging in the EModE period are mainly adverbs like *willingly* and *accordingly*, ... (2012b: 744–745)

examples of *accordingly*. A16–2 (1551–1600) shows four examples (all illustrating *as*) among the total of 50. Hereafter, relevant examples are scanty, both A17–1 (1601–1650) and A17–2 (1651–1700) providing only one example each.

19. Swan’s (1991: 409) definition of the adverbial shift runs as follows: “By an adverbial shift I mean the change of scope whereby an adverb comes to modify a different element, group of elements, etc. than it did initially, ... e.g. from being word-modifying to becoming sentence-modifying”.

It is true that *willingly* is quite frequently attested in Early Modern English. *Accordingly* is also common in the same period, appearing not in the writings by More but in Bacon, Baxter, and Locke as well as in EMEPS-A in my data. The above sections have, however, shown that the inventory of -ingly adverbs in Early Modern English is not as poor as assumed: SAEMEP and EMEPS-A combined together provide more than 70 different types. It is, therefore, worth examining whether they include at least some examples of Harry Potter adverbs.

The definition of Harry Potter adverbs is given in Section 1. Despite the fairly straightforward definition, however, the exact judgement of Harry Potter adverbs is not always easy. Broccias himself often uses the phrase "... is *probably* not the Harry Potter type" (emphasis mine) (2012a: 152), indicating how elusive this category is. I would still consider that there are fairly clear examples of Harry Potter adverbs already in the Early Modern English period. Some illustrative examples found in SAEMEP are seen in (29)–(31).

- (29) Clisters also helpe, lest the Medicine stop in the Guts, and worke *gripingly*.
(Bacon, 1627, *Sylua syluarum*)
- (30) Mr. Dodwell, *self-condemningly* saith, that a presumptuous Ordination of the Priest serves to the validity of Sacraments, though indeed he were not Ordained
(Baxter, 1682, *An Answer to Mr. Dodwell and Dr. Sherlocke*)
- (31) Many indeed who have jump'd in with the Hypothesis of Epicurus, have led Lives unworthy of his Patern, and spoke more *slightingly* of Matters which we ought to revere, than he did. (Locke, 1694, *Reason and Religion*)

As mentioned above, it is not an easy task to decide whether a particular example is of the Harry Potter type, which is partly due to the fact that the derivative process of some -ingly forms, namely whether the base is adjectival or verbal, is often difficult to tell. Still, I would state that there are about 15 relatively clear examples of Harry Potter adverbs, and about 20 if possible adverbs are included, in SAEMEP. Whereas this may look like a small quantity when viewed from the total of 350 occurrences of -ingly adverbs in this corpus, it is not necessarily the case. More than 300 out of the 350 examples are established and therefore common ones, that is *accordingly*, *cunningly*, *everlastingly*, *exceedingly*, *willingly*, *wittingly*, and *lovingly*. Since Broccias (2012a) cautiously excludes these examples, I have excluded them while counting the candidates. The likelihood for the remaining -ingly adverbs to be of the Harry Potter type is not low.

Plausible examples of Harry Potter adverbs are even more numerous in EMEPS-A, which includes texts written by various authors and which yields a larger inventory of -ingly adverbs to begin with. See, for instance, Examples (32)–(37).

- (32) For the lyuely voyce of God, *rebukyngly* toke me vp, before I wente out of the place. (Queen Marguerte, 1548, *A Godly Medytacyon of the Christen Sowle*)
- (33) This is no bread, but said *affirmingly*, This is my body. not denying the bread, but affirming that his body was eate~, (meaning spiritually) as the breade was eaten corporally. (Thomas Cranmer, 1550, *A Defence of the True and Catholike Doctrine of the Sacrament of the Body and Bloud of our Sauour Christ*)
- (34) vvhereby the nevv pulpits (the very chaires of the skorneful) merily or *mockingly* often call vs and our brethren. (William Allen, 1581, *An Apologie and True Declaration of the Institution and Endeuous of the Tvvo English Colleges*)
- (35) Parry neuerthelesse persisted still in his rage and fonde speach, and *ragingly* there sayd, he there su~moned Queene Elizabeth to answere for his blood before God (William Parry, 1585, *A True and Plaine Declaration of the Horrible Treasons*)
- (36) but on the contrary, boasting of his own innocence, *vauntingly* affirmed, even to the scandal of her Majesties honour, that though he knew a certain means to secure his life, yet he would rather suffer a thousand deaths, then not leave a testimony to the world, that he despised both her favour aud her power. (W. Sales, 1655, *Theophania*)
- (37) And when Julian Apostatized from the Faith, banished the Gospel, demolished the Churches, and *scoffingly* told the Christians, their Doctrine was to suffer (White Kennett, 1689, *A Dialogue between Two Friends Occasioned by the Late Revolution of Affairs*)

Even in the Middle English data, there are some possible examples, as the following citations ((38)–(40)) from METiP illustrate.

- (38) They answerde ful falsly and *flateryngly* & saide nay (Trevisa, 1482, *Polychronicon*)
- (39) Ne who most *felyngly* speketh of loue (Chaucer, 1477, *The Canterbury Tales*)
- (40) To whom she answerd *smylyngly* (Raymond of Capua, 1500, *Saint Katherin of Senis*)

I do not intend to deny Broccias's remark that Harry Potter adverbs are essentially products of Late Modern English, since it may well be the case that their notable expansion indeed took place only in the later period of Modern English. Still, such examples as quoted above demonstrate that the Late Middle and Early Modern English periods should not be ignored while discussing *-ingly* adverbs of this type.

6. Conclusion

The above discussion has dealt with the rise of *-ingly* adverbs in the Late Middle and Early Modern English periods. Not only the frequency but also the inventory of *-ingly* adverbs expanded during this period: in Late Middle English there were only a limited number of *-ingly* adverbs shared by a number of texts. Malory is most peculiar in his choice of *-ingly* adverbs, giving only two types, of which *wittingly* occurs only once. The use of *-ingly* adverbs in this text is virtually confined to *pass-ingly*, which functions as a modifier of adjectives or adverbs.

In the Early Modern English period, the inventory of *-ingly* adverbs expanded, especially after Thomas More. It is characteristic of this period that some limited number of *-ingly* adverbs were noticeably frequent, showing clearly that they were established, for instance *accordingly*, *willingly*, and *everlastingly*. Other items are, however, rather occasional, occurring only once or twice or in particular texts. The path of the lexicalization of *-ingly* adverbs, therefore, seems to be motivated by an individual's creativity in forming or adopting new forms.

The last section discussed the functions of *-ingly* adverbs in Late Middle and Early Modern English. Although most examples functioned to modify verbs or verb phrases in this period, there were also examples of other types. As mentioned above, Malory is quite unique in this respect: his use of *-ingly* adverbs is virtually confined to the modification of adjectives or adverbs. Otherwise, examples modifying adjectives locally were not common throughout the period explored. Likewise, the modification of the whole sentence was not common. This seems to be a late development: in my corpora, relevant examples are encountered only in the Early Modern English period, most representatively illustrated by *accordingly* used as a conjunct. The disjunct use of *-ingly* adverbs was not yet attestable even in Early Modern English. The adverb *seemingly*, which shows writers' or speakers' inference, is observed from the 17th century onwards, but it still modifies adjectives locally rather than sententially, in the data investigated in the present study. Hence, it is safe to state that *-ingly* adverbs in Early Modern English are still at a stage prior to the development of the all-round functions as found in Present-day English.

Finally, the above discussion has dealt with what Broccias (2012a) calls Harry Potter adverbs. According to him, *-ingly* adverbs of this type are really attested from the Late Modern English period onwards. It is true that Harry Potter adverbs are not exactly numerous in the Early Modern English period, but the above discussion has shown that they can already be encountered in the period investigated in the present study. They are available even in METiP, which includes 15th-century printed materials.

Acknowledgements

This study was in part supported by JSPS Kakenhi (Grant Number 18K00645). I am grateful to the two anonymous reviewers and the editors of this volume for a number of comments and suggestions on the first draft of this paper.

References

Corpora

EEBO = *Early English Books Online*. <<http://eebo.chadwyck.com/home>> (1 November 2018).

Other references

- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Broccias, Cristiano. 2012a. Oriented *-ingly* adjuncts in Late Modern English. In *English Historical Linguistics 2008, Vol. 2: Words, Texts and Genres* [Current Issues in Linguistic Theory 324], Hans Sauer & Gaby Waxenberger (eds), 147–164. Amsterdam: John Benjamins.
- Broccias, Cristiano. 2012b. The syntax-lexicon continuum. In *The Oxford Handbook of the History of English*, Terttu Nevalainen & Elizabeth Closs Traugott (eds), 735–747. Oxford: OUP.
- Iyeiri, Yoko. 2011. Early Modern English Prose Selections: Directions in historical corpus linguistics. *Memoirs of the Faculty of Letters, Kyoto University* 50: 133–199. <https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/139204/1/lit50_133.pdf> (8 December 2018).
- Iyeiri, Yoko. 2014. The shift from *alway* to *always* in the history of English. In *Studies in Middle and Modern English: Historical Change*, Yoko Iyeiri & Jennifer Smith (eds), 29–47. Suita: Osaka Books.
- Iyeiri, Yoko. 2016. On *alway(s)* and *algate(s)* in Middle English again. *Memoirs of the Faculty of Letters, Kyoto University* 55: 61–80. <https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/209963/1/lit55_61-80.pdf> (8 December 2018).
- Ker, N. R. (ed.) 1976. *The Winchester Malory: A Facsimile*. London: OUP.
- Killie, Kristin. 1998. The spread of *-ly* to present participles. In *Advances in English Historical Linguistics (1996)*, Jacek Fisiak & Marcin Krygier (eds), 119–134. Berlin: Mouton de Gruyter.
- Killie, Kristin. 2000. *Stative Adverbs in English: A Study of Adverbial Productivity and Orientation*. PhD dissertation, the University of Tromsø.
- Kjellmer, Göran. 1984. Why *great*: *greatly* but not *big*: **bigly*: On the formation of English adverbs in *-ly*. *Studia Linguistica* 38: 1–19.
- Middle English Dictionary (MED)*. <<https://quod.lib.umich.edu/m/middle-english-dictionary>> (16 February 2019).
- Nevalainen, Terttu. 1994. Aspects of adverbial change in Early Modern English. In *Studies in Early Modern English*, Dieter Kastovsky (ed.), 243–259. Berlin: Mouton de Gruyter.

- The Oxford English Dictionary* on CD-ROM. 1989. 2nd edn. John A. Simpson & Edmund S. C. Weiner (eds). Oxford: OUP.
- Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Raumolin-Brunberg, Helena & Nevalainen, Terttu. 1994. Social conditioning and diachronic language change. In *Studies in Early Modern English*, Dieter Kastovsky (ed.), 325–338. Berlin: Mouton de Gruyter.
- Rowling, J. K. 2000. *Harry Potter and the Goblet of Fire*. London: Bloomsbury.
- Swan, Toril. 1991. Adverbial shifts: Evidence from Norwegian and English. In *Historical English Syntax*, Dieter Kastovsky (ed.), 409–438. Berlin: Mouton de Gruyter.
- Swan, Toril. 1997. From manner to subject modification: Adverbialization in English. *Nordic Journal of Linguistics* 20: 179–195. <https://doi.org/10.1017/S0332586500004108>
- Swan, Toril. 2003. Present participles in the history of English and Norwegian. *Neuphilologische Mitteilungen* 104: 179–195.

Appendix

A detailed list of the texts included in Table 2

Authors	Texts (short titles)
Sir Thomas More (1478–1535)	1529 <i>The supplycacyon of soulys</i> 1529 <i>A dyaloge of syr Thomas More knyghte</i> 1533 <i>A letter of syr Tho. More knight</i> 1533 <i>The answeere to the fyrst parte of the poysened booke, which a namelesse heretyke hath named the souper of the lorde</i> 1533 <i>The apologye of syr Thomas More knight</i> 1553 <i>A dialoge of comfort against tribulacion, made by Syr Thomas More Knyght</i>
Francis Bacon (1561–1626)	1601 <i>A declaration of the practises & treasons attempted and committed by Robert late Earle of Essex and his complices</i> 1603 <i>A briefe discourse</i> 1604 <i>Sir Francis Bacon his apologie</i> 1614 <i>The charge of Sir Francis Bacon Knight</i> 1618 <i>A declaration of the demeanor and cariage of Sir Walter Raleigh, Knight</i> 1629 <i>The historie of the reigne of King Henry the Seuenth</i> 1638 <i>The historie of life and death</i> 1642 <i>The learned reading of Sir Francis Bacon</i> 1657 <i>Resuscitatio, or, Bringing into publick light severall pieces of the works, civil, historical, philosophical, & theological, hitherto sleeping, of the Right Honourable Francis Bacon</i> 1627 <i>Sylua sylvarum: or A naturall historie in ten centuries</i>

Authors	Texts (short titles)
Richard Baxter (1615–1691)	1655 <i>Aphorismes of justification</i>
	1655 <i>The arrogancy of reason against divine revelations</i>
	1658 <i>A call to the unconverted to turn and live and accept of mercy</i>
	1660 <i>Catholick vnity</i>
	1660 <i>The Christian religion expressed</i>
	1661 <i>An accompt of all the proceedings of the commissioners of both persvvasions appointed by His Sacred Majesty</i>
	1682 <i>Additional notes on the life and death of Sir Matthew Hale</i>
	1682 <i>An answer to Mr. Dodwell and Dr. Sherlocke</i>
	1689 <i>Cain and Abel malignity</i>
	1691 <i>Church concord</i>
	1693 <i>The Christians converse with God</i>
John Locke (1632–1704)	1670 <i>The fundamental constitutions of Carolina</i>
	1690 <i>An essay concerning humane understanding microform</i>
	1690 <i>Two treatises of government in the former</i>
	1692 <i>Some considerations of the consequences of the lowering of interest</i>
	1694 <i>Reason and religion</i>
	1695 <i>A vindication of The reasonableness of Christianity, &c</i>
	1697 <i>A second vindication of The reasonableness of Christianity, &c</i>

Funding information

Funded by: Japan Society for the Promotion of Science, Kakenki.

Award ID: 18K00645.

Analyzing change in the American English amplifier system in the fiction genre

Martin Schweinberger
University of Queensland

This study examines the diachronic development of amplification of adjectives in American English (AmE), specifically in the fiction section of the *Corpus of Historical American English* (COHA). The results show that amplifier use in attributive contexts remains stable, while, in predicative contexts, *so* has increased significantly since the 1980s and has replaced *very* as the dominant amplifier. This development contrasts with ongoing change in other regional varieties of English where *really*, rather than *so*, has become the dominant amplifier of adjectives during the latter half of the 20th century. The results show that *so* and *very* are semantically highly similar and that the increase in *so* is mostly based on its co-occurrence with *good*.

1. Introduction

This study takes a corpus-based approach using the fiction section of the *Corpus of Historical American English* (hereafter COHA, Davies 2010) to investigate long-term, but ongoing, changes in the amplification of adjectives in American English (see the examples in (1)). Amplification is a subtype of intensification, which is related to the semantic category of degree. For this reason, adverbs that serve to intensify adjectives have also been referred to as degree adverbs or adverbs of degree (Quirk et al. 1985). Intensification ranges between very low intensity (downtoning) and very high intensity (amplifiers) (Quirk et al. 1985: 589–590). According to Quirk et al. (1985: 589–590), amplifiers “scale upwards from an assumed norm [while] downtoners have a lowering effect, usually scaling downwards from an assumed norm” (1985: 590). The current chapter exclusively focuses on amplifiers, while leaving aside downtoners (which include approximators such as *almost*, compromisers such as *more or less*, diminishers such as *partly*, and minimizers such as *hardly*). Within the category of amplifiers, Quirk et al. (1985: 589–590) differentiate between maximizers such as *completely*, which denote the upper extreme of the

scale (Quirk et al. 1985: 590) and boosters such as *very*, which denote a high degree or a high point on the scale. Boosters, in particular, form an open class, which adopts new members to replace forms which have lost their expressiveness due to frequent use (see Quirk et al. 1985: 590).

- (1) a. I feel assured that I looked *very* terrific. (COHA: fic_1846_200)¹
- b. You're *pretty* nice after all. (COHA: fic_2001_46155)
- c. That is *so* awesome. (COHA: fic_2007_34397)
- d. It's not too long, and it's *so* cool. (COHA: fic_1980_10659)
- e. Your grandma taught me years ago to make *very* good French toast.
(COHA: fic_1995_39423)
- f. ...whenever they completed a *really* good job. (COHA: fic_1966_10471)

From a language variation and change perspective, intensifiers and amplifiers are particularly interesting, as intensifier systems are prone to change (Brinton & Arnovik 2006: 441; Ito & Tagliamonte 2003: 257; Quirk et al. 1985: 590). For this reason, intensification has been referred to as a domain of “fevered invention” (Bolinger 1972: 18) and an area of waxing and waning of forms, alongside invention and renewal (D’Arcy 2015: 450). The latter is particularly intriguing as the changing nature predestines intensifier systems to be an ideal case for testing mechanisms of change such as determining conditioning factors of lexical replacement, which is the topic of this chapter. From the point of view of pragmatics, amplifiers are intriguing because they play a crucial part in how speakers express themselves socially and emotionally (Ito & Tagliamonte 2003: 258; Labov 1985: 43). Thus, intensifiers are part of an inventory on which speakers rely to create and mark their social identity (Tagliamonte 2011: 30).

Although there already exists a substantial body of research on intensification, the present study adds to existing research. It investigates the subsystem of amplification of adjectives by considering all variants that occur in the variable context, rather than analyzing individual or only a small set of selected amplifiers. Furthermore, the current analysis devotes itself specifically to the interdependencies between amplifiers and adjectives. Recent research has shown that adjectives play a critical role in facilitating changes in amplifier use (see Tagliamonte 2011: 144; Wagner 2017). In addition, the current study extends existing research by relying on a dataset of substantial size which renders the current analysis interesting as

1. The annotation of the examples indicates first the corpus (COHA), followed by the genre (*fic* indicating fiction), then the time of publication (1846 therefore shows that the respective text was published in 1846), and finally the text number (200 stands for text 200 of the fiction section published in 1846).

relying on a larger dataset bears the potential of detecting changes in collocational preferences that would not be detectable given datasets of more moderate size. In terms of methodology, the current study adds to previous research by zeroing in on a well-circumscribed and restrictive variable context (similar approaches have been undertaken by D'Arcy 2015; Tagliamonte 2008; Tagliamonte & Denis 2014; Tagliamonte & Roberts 2005), namely, the variation of amplifier variants in pre-adjectival slots. Restricting the analysis to a narrowly defined variable context allows the precise evaluation of factors contributing to variation, which would not be possible if the analysis built on relative frequencies (see Labov 1966: 49; Tagliamonte 2011: 9–10). The aim of the current study is to add to our understanding of mechanisms that underly lexical replacement which is one of the core mechanisms of language change.

Before turning to an overview and discussion of previous research on amplifiers, a note on the historicity of amplifier variants and terminology is in order. Although variants which become dominant are referred to here as innovative variants, the term 'innovative' does not refer to a recent origin but to changes in their linguistic behavior. Indeed, most – if not all – amplifiers which become highly frequent have a long-standing history. The most prominent amplifiers in present-day varieties (*very*, *really*, and *so*) have been in use as amplifiers for centuries. *So* – or its original form *swa* – has even been used as in amplifier in *Beowulf*, written between 975 and 1010 (see the instances provided in Tagliamonte & Roberts 2005: 293). Whereas *very* and *really* originally started out as adjectives during Middle English and conveyed senses of 'true' and 'real' respectively, *very* has become prominent since the 16th and *really* during the 18th century (Mustanoja 1960: 323–324; see also Ito & Tagliamonte 2003: 260). In other words, innovative variants are not new, but their behavior is innovative in the sense that they undergo a functional or collocational change and thereby undergo a transition in terms of use.

The next section presents previous research on amplifiers and changes in amplifier systems. Section 3 introduces the data used in the current study and elaborates on the steps undertaken during data processing. Section 4 provides the results of the statistical analysis and Section 5 discusses the results in light of previous research and evaluates shortcomings of the present analysis while Section 6 offers an outlook for future research.

2. Previous research

Intensification has been subject to linguistic analyses for more than a century and various studies have been dedicated to investigate degree modification from descriptive perspectives (e.g. König 2017; Paradis 2008) but also from a historical or language change perspective (see, for instance, Bolinger 1972; Borst 1902; Breban & Davidse 2016; Fettig 1934; Lorenz 2002; Méndez-Naya 2003, 2008; Méndez-Naya & Pahta 2010; Nevalainen 2008; Nevalainen & Rissanen 2002; Partington 1993; Peters 1992, 1993, 1994; Rissanen 2008; Stoffel 1901). Indeed, given the substantial amount of attention that intensification has attracted, the developmental pathways of individual or selected intensifiers has been subject to intensive research (e.g. Aijmer 2011, 2018a; Macaulay 2006; Méndez-Naya 2003; Núñez-Perotejo & Palacios Martínez 2014; Rickford et al. 2007; Tao 2007). Also, the use of amplifiers which intensify adjectives by native and non-native English speakers has recently attracted attention (Fuchs 2016). Furthermore, frequency differences in the usage of adjective amplifiers based on the age, gender, or the social class of speakers have undergone a good amount of scrutiny (e.g. Bauer & Bauer 2002; Fuchs 2017; Ito & Tagliamonte 2003; Macaulay 2002; Núñez-Perotejo & Palacios Martínez 2014; Palacios Martínez & Núñez-Perotejo 2012; Stenström 1999; Stoffel 1901; Tagliamonte 2008; Tagliamonte & Denis 2014; Tagliamonte & Roberts 2005). The findings of these sociolinguistic studies are detailed below. In addition, the distribution of amplifiers across registers (see, for example, Fuchs 2016) and their use across varieties of English are well understood (see, for instance, Aijmer 2018b; Bauer & Bauer 2002; Calle-Martín 2014; Fuchs 2016; de Klerk 2005; Paradis 1997; Tagliamonte & Denis 2014; Wagner 2017; Xiao & Tao 2007). Also, various fine-grained variationist studies have provided detailed analyses of the trajectories of ongoing changes in intensifier systems of various varieties of English (e.g. D'Arcy 2015; Ito & Tagliamonte 2003; Tagliamonte 2008; Tagliamonte & Denis 2014; Tagliamonte & Roberts 2005). These sociolinguistic and variationist studies that have analyzed the use of intensifiers within and across varieties of English have detected a common trend in that *very* is declining in both apparent and real time while *really* is increasing (see D'Arcy 2015 for New Zealand English (NZE); Ito & Tagliamonte 2003 and Barnfield & Buchstaller 2010 for North East British English; Tagliamonte 2008 and Tagliamonte & Denis 2014 for Toronto English; and Tagliamonte & Denis 2014 for South Eastern Ontario English). Due to the reciprocal nature of the waxing and waning of forms, previous studies (e.g. D'Arcy 2015) have suggested that these developments represent cases of a particular long-term trend of lexical replacement. Furthermore, these studies showed that re-arrangements in amplifier systems, for instance, the partial lexical replacement of *very* by *really*, are accompanied by ordered extra- as well as

intra-linguistic stratification (e.g. D'Arcy 2015; Lorenz 2002; Tagliamonte 2008; Tagliamonte & Roberts 2005). Before turning to a more detailed discussion of the findings presented in previous research, it should be mentioned that while adverbs of degree have been studied intensively, remarkably little research has focused on genre differences and, in particular, their use in written genres. One notable exemption is Biber et al. (2007) who briefly discuss differences in the use of amplifiers in British and American English conversation as well as academic discourse. Regarding differences between spoken conversational and written academic texts, Biber et al. (2007: 565) found that *entirely*, *extremely*, *fully*, *highly*, and *strongly* are used more frequently in academic discourse, while the use of *very* was remarkably stable across registers. In addition, amplifiers had fewer collocations in academic prose compared with spoken conversational data.

With respect to intra-linguistic stratification and language internal constraints on the replacement of a dominant form by a more innovative variant, syntactic function is among the most consistent factors (see Mustanoja 1960: 326–327 and Tagliamonte 2008: 373). Collocation with adjectives in predicative function is commonly regarded to indicate later stages of change (see Tagliamonte 2008: 363 and Tagliamonte & Denis 2014: 116). In initial stages of change, innovative variants are typically associated with attributive contexts (Mustanoja 1960: 326–327; Tagliamonte & Denis 2014: 116) and only later enter predicative contexts (D'Arcy 2015: 471–472). This patterning is caused by a general process of grammaticalization and more specifically, delexicalization. During the delexicalization process, forms enter syntactic contexts that they were previously barred from entering due to having lost semantic content (semantic bleaching/delexicalization).

One of the best attested clines for delexicalization is the grammaticalization of *very* from an adjective to intensifying adverb. The delexicalization of *very* proceeded in four steps. Initially, *very* functioned as an adjective meaning 'genuine, true'. In the next stage, it was used in coordinate constructions with a following attributive adjective (see Tagliamonte 2008: 363). At this stage, the demarcation between adjective and adverb is difficult, if not impossible, as the syntactic context does not allow for an unambiguous interpretation. During the last two stages, *very* is used as an intensifier first with attributive and only then with predicative adjectives. Also, at this stage the semantic meaning of *very*, 'genuine, true', has been entirely lost, leaving solely the intensifying function. In analogy to the trajectory of *very*, modification of predicative adjectives has been deemed a sign of advanced delexicalization (Tagliamonte 2008: 373) and it appears to be a general trend that innovative variants occur with attributive adjectives first (Mustanoja 1960: 326–327). Only during later stages do innovative forms occur with predicative functions (Tagliamonte 2008: 363).

Another factor which frequently accompanies changes in intensifier use is the association of innovative variants with negative polarity items or emotional adjectives more generally (see, for example, Peters 1994). This, of course, applies predominately to elements which are recruited from negative domains but not elements which are recruited from a positive domain such as *very*. Indeed, Partington (1993: 184) argues, in this respect, that the reason for the correlation between innovative intensifiers and (negative) emotional adjectives is grounded directly in the negative domain from which intensifiers are often recruited (*terrible, horrible*, etc.). According to Partington (1993: 184), these negative items must first undergo delexicalization before modifying positive polarity items such as *good* or *nice*. In other words, during early stages of change, intensifiers tend to associate with semantically negative adjectives and only during later stages, when the incoming forms have been bleached, do they collocate with positive adjectives as semantic restrictions wane. With respect to ongoing changes, Tagliamonte and Roberts (2005: 289) found that the innovative intensifier *so* significantly collocates with emotional adjectives (particularly among female speakers), as does *really* among speakers between the ages of 20 and 29 in Toronto (Tagliamonte 2008: 383).

One of the most intriguing aspects that has been described in the respective literature relates to collocation preferences of intensifiers. Innovative forms are restricted to a relatively small and fixed set of adjectives and this set only expands once a form becomes more frequent and as collocational restrictions erode (see Lorenz 2002: 144; Méndez-Naya 2003: 377; and Tagliamonte 2008: 376). In this respect, Tagliamonte and Roberts (2005) show that *really* collocates with high-frequency adjectives (and thus presumably correlates positively with adjective frequency) whereas *so* – as the most innovative intensifier in their data – prefers adjectives that are neither particularly frequent nor infrequent. The traditional form *very*, however, is the most general form with respect to its distribution and only shows a marked underuse for the three most frequent adjectives in the data of Tagliamonte & Roberts (2005). In addition, Tagliamonte and Denis (2014) show, in their research on the trajectory by which innovative variants come to dominate a linguistic subsystem, that successful variants first associate with few forms and only expand across semantic classes of adjectives, that is, show an increase in their lexical diversity, after they have become well established (Tagliamonte & Denis 2014).

Concerning extra-linguistic stratification, the situation is somewhat complex. Apparent-time distributions as well as multivariate statistics confirm consistent trends for age – with younger speakers preferring *really* and other innovative forms such as *dead*, *pretty* or *so* (see Bauer & Bauer 2002 for an interesting study on boosters in NZE) while older speakers exhibit a marked preference for *very* (D'Arcy 2015; Ito & Tagliamonte 2003; Tagliamonte 2008) – whereas the results for gender,

in particular, are less coherent. Fuchs (2017) shows that in British English, women, overall, use more amplifiers compared with men and that this gender difference remains rather stable across real time. Concerning individual amplifiers, Bauer and Bauer (2002), as well as Tagliamonte and D'Arcy (2009), found significant but weak gender differences. Tagliamonte and D'Arcy (2009) found minor gender differences in the use of amplifying *so*, and Macaulay (2006) found a female lead for *pure* as an intensifier in Glasgow English. Ito and Tagliamonte (2003) as well as Tagliamonte (2008) did not find a consistent effect of gender, as gender differences were restricted and dependent upon speaker age, which shows that "the use of intensifiers by male and female speakers of different ages is intimately tied to the stages of intensifier renewal in the community grammar" (Tagliamonte 2008: 385). To complicate matters further, D'Arcy (2015: 477) found that ongoing change in the amplifier system did not show any significant gender differences in NZE.

The next section provides an overview of the data the current study is based on and elaborates on matters of data processing.

3. Data and methodology

The current study draws on data from the fiction section of the *Corpus of Historical American English* (COHA, Davies 2010). The COHA is part of the Brigham Young University (BYU) family of corpora created by Mark Davies and it represents the largest structured corpus of historical English comprising approximately 385 million words (see the 'overview' section at <https://www.corpusdata.org>). The current study has used the downloaded full-text version of the COHA where 10 words of every 200 words are removed and replaced with "@" (see the 'limitations' section at <https://www.corpusdata.org>). The COHA comprises texts published between 1810 and 2009 from the genres non-fiction, fiction, magazines, and news. Only fiction texts are considered in the present study. It should be borne in mind in this context that the fiction genre is more speech-like compared with the other text types represented in the COHA as fiction is likely to incorporate text that aims to resemble spoken discourse. As such, the decision to concentrate exclusively on the fiction genre is motivated by its relative speech-like character.

To extract all adjectives, the corpus data was part-of-speech tagged in the programming environment R (R Development Core Team 2008) by implementing a maximum entropy tagger provided in the openNLP package (Hornik 2016). After part-of-speech-tagging, all adjectives (part-of-speech tag "JJ") were extracted – this encompassed base forms, including participle forms, but excluded comparative and superlative forms. This means that the analysis exclusively focuses on

adjective modification. Next, it was determined for each adjective whether it was amplified based on a predefined set of amplifiers. Adjective tokens that were modified by *too* were excluded as *too* implies an inherent but oftentimes not expressed comparison. The next step was to add the decade in which the text was produced and the syntactic function of the adjective (predicative vs. attributive) was determined. If the syntactic function could not be determined, the adjective was removed from the analysis.

Furthermore, if an adjective token was preceded by a downtoning intensifier (e.g. *sort of, kind of, a little bit, somewhat, slightly, barely, hardly*, etc.), by a negative form (e.g. *not, never*, etc.), by forms indicating comparison (*more* and *most*), or by a special form (i.e. *too, quite*, etc.), the respective adjective token was removed from the analysis. Next, adjective types that were never amplified or that were not amplified by at least two distinct amplifier types were removed from the analysis to prevent the inclusion of lexicalized forms (e.g. *right honorable*) and to allow zooming in on a well-circumscribed, restrictive variable context. Then, the data was manually cross-evaluated and cross-tabulated. Cross-tabulation of the data showed that the data before the 1850s should be ignored, due to the relatively moderate amount of data points available for this period. Also, it exhibited idiosyncratic patterns with respect to amplifier-adjective bigrams which were likely resulting from the data not being balanced to a degree that would be necessary to warrant robust results.

Next, a semantic vector space model based on the co-occurrence frequencies of amplifiers and adjectives was applied to the data. The implementation of the semantic vector space model followed Levshina (2015) and resulted in the dendrogram presented in Figure 1. The amplifiers in red (i.e. *absolutely, entirely, totally, utterly, wholly*) were removed from the analysis, as the semantic vector space model determined that their collocational profile was significantly different from the profile of other amplifier types. The reasoning for excluding these variants is grounded in the principle of accountability (see Labov 1966: 49; Tagliamonte 2011: 9–10). For elements to be considered variants, they must be interchangeable. If the collocational profile is significantly different, then these elements do not represent variants as they are not interchangeable. This is somewhat problematic, as it limits the analysis to a subset of amplifiers (boosters) whereas the aim of the study was to investigate the subsystem of adjectival amplification as whole. However, including elements with significantly distinct collocational profiles would violate the principle of accountability.

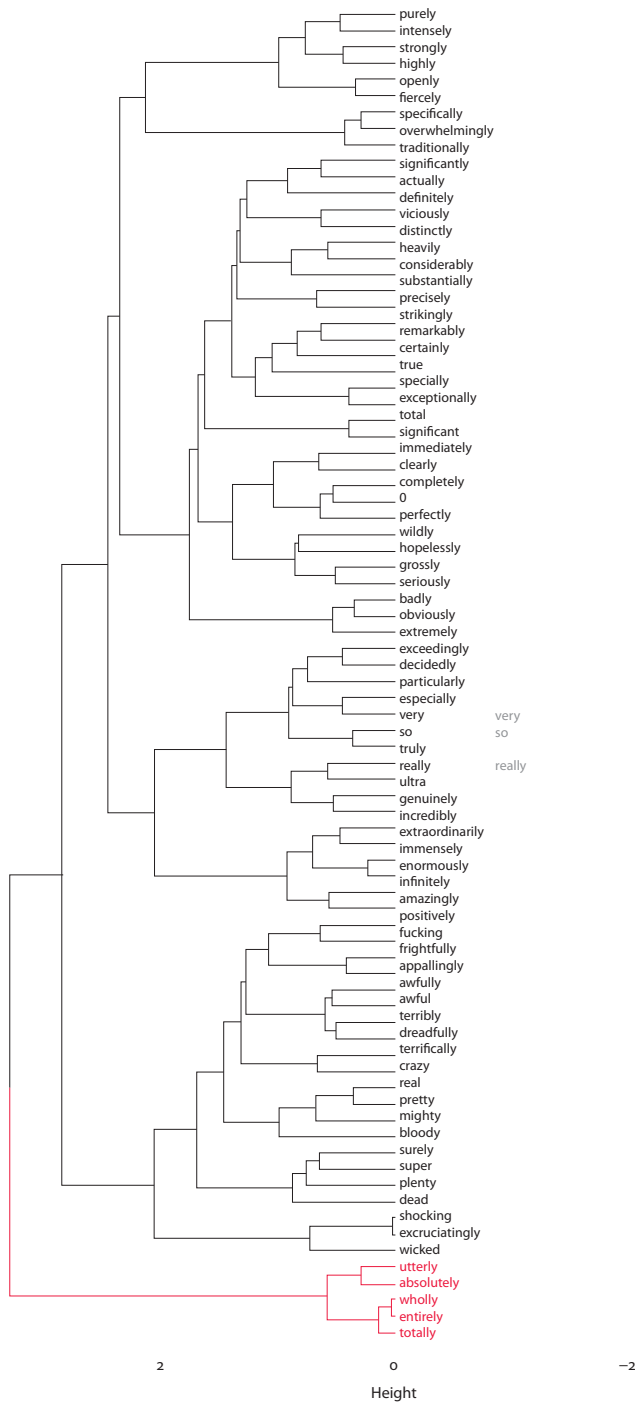


Figure 1. Horizontal cluster dendrogram showing the semantic similarity of amplifiers in COHA based on their collocational profile with adjectives

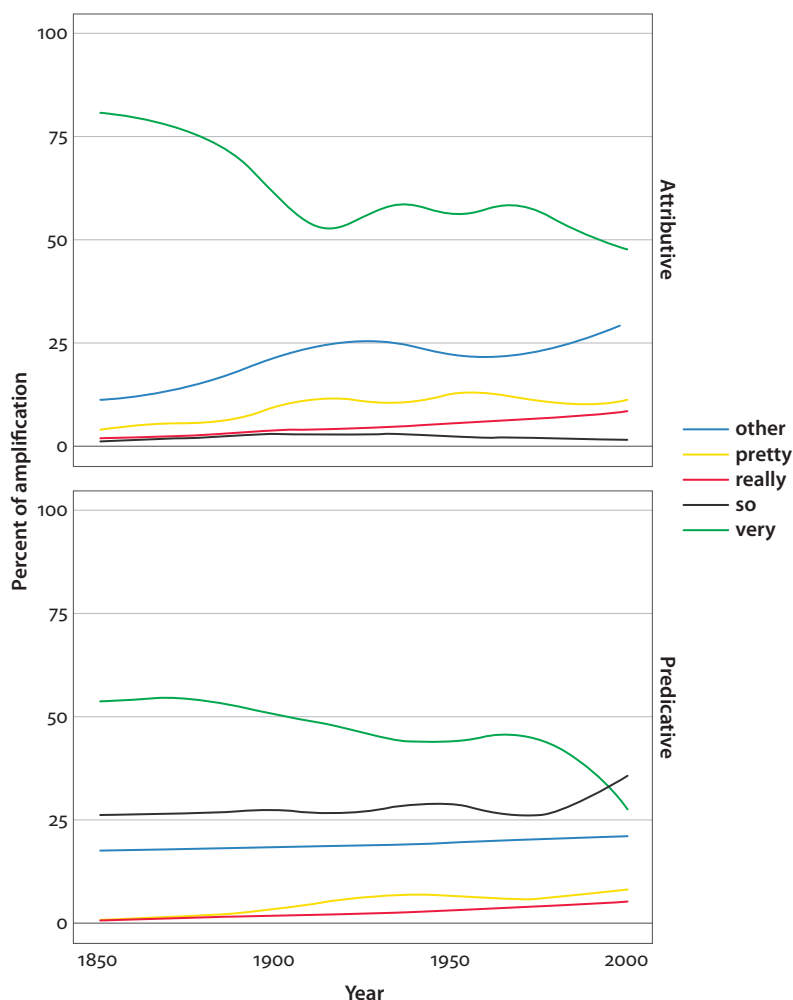


Figure 2. Distribution of amplifiers in attributive and predicative contexts in the fiction genre of the COHA across real time

To test if the American data indicate lexical replacement in terms of frequency, Figure 2 displays the distribution of amplifier types in attributive versus predicative contexts in the fiction genre of the COHA data across real time. According to Figure 2, *so* has replaced *very* as the dominant form in predicative contexts during the latter half of the 20th century. Since *so* – which appears to be the rival variant to *very* in the American fiction data – is almost categorically restricted to predicative contexts and because the distribution of amplifier types in attributive contexts is

rather stable,² the subsequent analysis disregards all attributive contexts and focuses on predicative contexts only.

After processing the data according to the steps described above, the remaining data consisted of 31,527 adjectives of which 19.63% were amplified. The most frequent amplifier in the data is *very* with 2,801 instances comprising 45.27% of all instances of amplification. The second most frequent amplifier is *so* with 1,747 instances representing 28.23% of overall amplification.

Table 1. Overview of the cleaned and processed data*

Variant	N	%	% (of Amp.)
Ø (not amplified)	25,339	80.21	
<i>very</i>	2,801	8.87	44.81
<i>so</i>	1,747	5.53	27.95
<i>pretty</i>	311	0.98	4.98
<i>really</i>	165	0.52	2.64
<i>real</i>	135	0.43	2.16
<i>extremely</i>	107	0.34	1.71
11–100: <i>perfectly, terribly, awfully, completely, mighty, highly, particularly, exceedingly, truly, dead, awful, especially, decidedly, distinctly, dreadfully, positively, seriously, incredibly, obviously, remarkably, genuinely, clearly, frightfully, intensely, certainly, extraordinarily</i>	869	2.77	13.93
1–10: <i>actually, immensely, infinitely, profoundly, amazingly, bloody, enormously, hopelessly, fucking, wildly, strikingly, definitely, exceptionally, greatly, plenty, immediately, specially, considerably, openly, surely, crazy, excruciatingly, heavily, overwhelmingly, purely, strongly, super, terrifically</i>	116	0.35	1.87
Total	31,590 (6,251)	100.0	100.1

* In Table 1, the amplifiers that are collapsed into frequency categories (1 to 10 and 11 to 100) are listed in the order of their token frequency.

Table 1 shows that *very* and *so* alone account for almost three quarters of all instances of amplification (72.76%). The vast majority of amplifiers occurs between 1 and 100 times and they account for merely 15.8% of all instances of amplification. We will now turn to the results of the current analysis.

2. While there is a notable decrease in *very* and an increase in the use of infrequent variants to compensate the waning of *very*, the overall ranking of variants remains identical across real time.

4. Results

The most general result relates to the percentage of amplified adjectives in variable contexts across real time. Displaying overall amplification is warranted here to render general trends in amplification visible (see Figure 3).

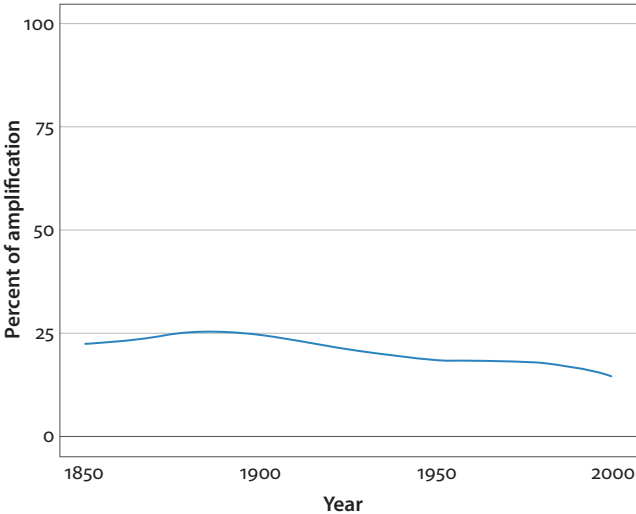


Figure 3. Percent of amplified adjective slots in predicative contexts in the fiction genre of the COHA after data processing (variable context)

Figure 3 shows the percentage of amplified adjective slots in predicative contexts in the fiction genre of the COHA across real time after data processing. The percentage values are restricted to the variable context, that is, not all adjective slots are considered. According to Figure 3, amplification in general appears to be in decline as more recent data shows lower percentage values of amplification compared with older data. Figure 4 shows the distribution of individual amplifier types across real time to offer a more fine-grained analysis.

Figure 4 shows that *so*, as the upcoming variant, has replaced the traditional *very* as the dominant form in real time in predicative contexts in fiction in American English since the 1990s. Apart from *very* and *so*, all other variants in the data remain remarkably stable and do not show substantial variation in their usage.

A second vector space model is now applied to a subset of the data in which non-amplified instances of adjectives are removed. The reason for implementing a second vector space model is to evaluate whether *so*, as the innovative or upcoming variant, is semantically similar to the dominant variant it replaces (*very*). The

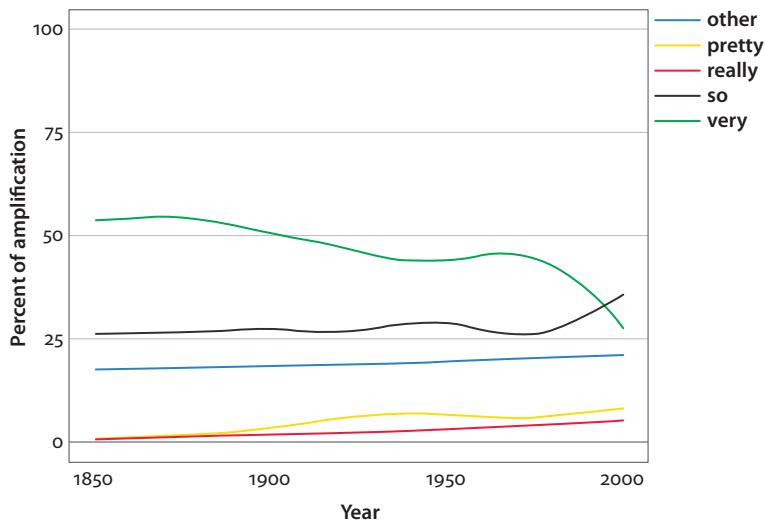


Figure 4. Distribution of amplifiers in predicative contexts in the fiction genre of the COHA across real time

vector space model allows for the inspection of the relative similarity of amplifier types based on their co-occurrence frequencies with adjectives (also referred to as vectors or embeddings). The underlying reasoning is founded in distributional semantics, according to which words that share many collocates are semantically similar (Stefanowitsch 2010: 368–370). In this view, semantic similarity can be measured in terms of the similarity of collocation profiles. In contrast to the previous semantic vector space model, this model concentrates explicitly on the very narrowly defined variable context and disregards cases in which an adjective is not amplified. Leaving out non-amplified adjectives is justified in this case because once a speaker has already decided to amplify, they then select a specific variant from a set of potential candidates. The vector space model is applied to test if distributional similarity affects the selection of a variant. If semantic similarity does indeed impact lexical replacement, then we should expect that (i) *so* and *very* are grouped closely together by the semantic vector space model and that (ii) the collocational profiles of *very* and *so* correlate very highly. Figure 5 shows the results of the second semantic vector space model as a dendrogram.

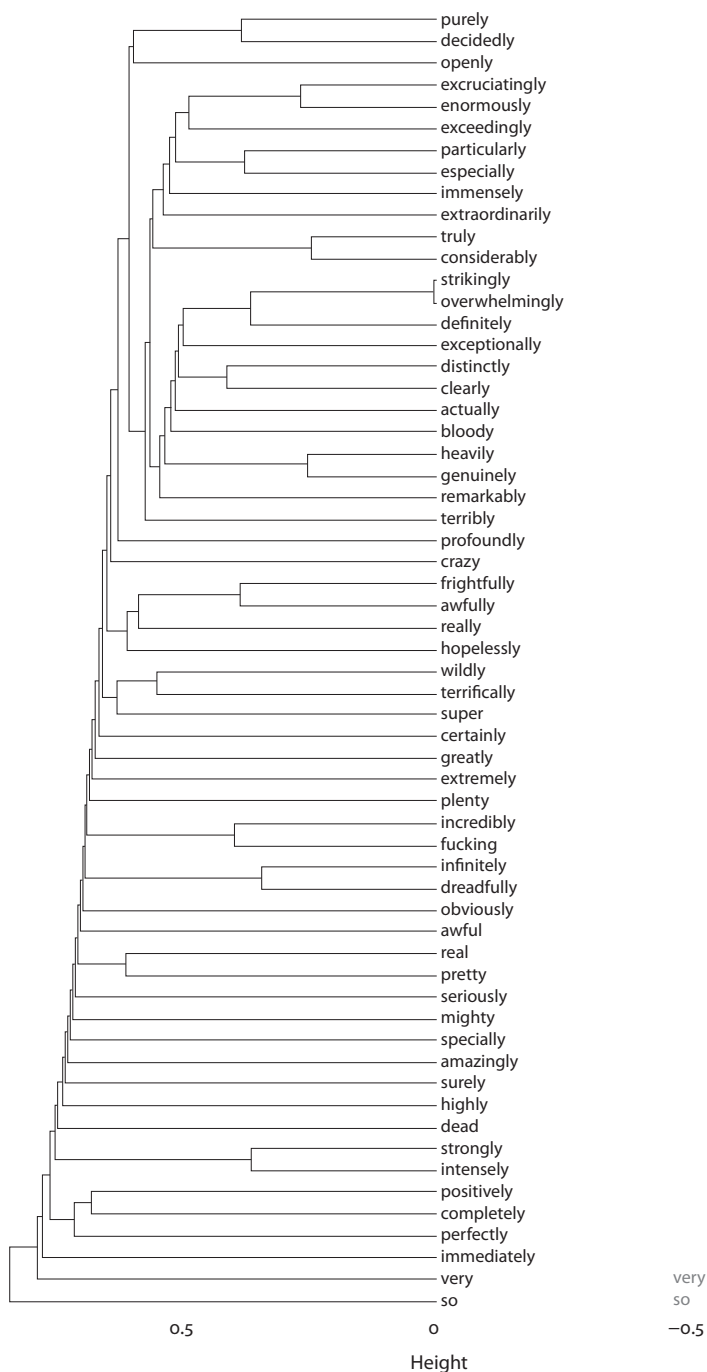


Figure 5. Dendrogram showing the results of a second vector space model applied to a subset of the data without regard for non-amplified adjectives

According to Figure 5, *so* and *very* are extraordinarily similar as they are separated by only a single node and correlate very strongly ($r = .68$).

Next, the analysis focuses on the lexical diversity of amplifier variants to provide insights into range of the collocational profile of variants. There are at least two potential pathways for a variant to become dominant. On the one hand, a variant could expand its collocational profile first and then show an increase in usage. This would mean that an increase in lexical diversity would be visible before there is an increase in the token frequency of an upcoming variant (high lexical diversity > increase in frequency of use). This trajectory would align with Mair's (2004) "delayed increase of discourse frequency" hypothesis (DIDFH) according to which increases in frequency appear after a form enters new functional contexts (see also Tagliamonte & Denis 2014: 117). On the other hand, a variant could 'specialize' first, that is, collocate only with a very small set of adjectives, before increasing in usage. An expansion of its collocational profile would then occur after the form has already become highly frequent. The second pathway would only be possible, of course, if the adjectives that the innovative variant associates with were high frequency adjectives, as only then could the innovative variant also experience a boost in usage. If this second pathway were correct, then the data should show an increase in frequency before the upcoming variant broadens its collocational profile, that is, co-occurs with a more diverse set of adjectives (increase in frequency of use > high lexical diversity). This second pathway is more likely to apply to amplification, as previous research (Tagliamonte & Denis 2014; Tagliamonte & Roberts 2005), has indicated that increases in frequency are accompanied by semantic expansion.

To analyze potential changes in lexical diversity, the current study calculates a simple lexical diversity measure (LDM) by dividing the number of distinct adjective types that an amplifier co-occurs with by the number of tokens of the respective amplifier (see formula in (2)).

$$(2) \quad LDM = \frac{N(\text{adjective types})}{N(\text{amplifier tokens})}$$

According to the formula in (2), the higher the LDM value, the higher the lexical diversity of a given amplifier. For example, if an amplifier occurs ten times but only co-occurs with a single adjective type, then this amplifier would have a lexical diversity measure LDM of $1/10 = 0.1$. If an amplifier occurs ten times and co-occurs with five different adjective types, it will have an LDM of: $5/10 = 0.5$. If an amplifier occurs ten times and modifies a different adjective type in each of these ten cases, it will have an LDM of $10/10 = 1$. The LDM values are calculated for *very*, *so*, and all other amplifiers combined for each decade in the dataset. All amplifiers except *very* and *so* had to be collapsed because *very* and *so* are the only amplifiers that occur in all decades.

The resulting distribution of LDM values across real time is shown in Figure 6. Figure 6 displays a consistent and rather stable patterning. The lexical diversity of amplifiers decreases across real time and *very* consistently has the lowest lexical diversity value. *So* has a lower lexical diversity compared with all other amplifiers combined, but its value is consistently higher than that of *very*. Interestingly, there is no noticeable change in the patterning between the 1970s and the 2000s, that is, the time period during which *so* replaced *very* as the dominant variant.

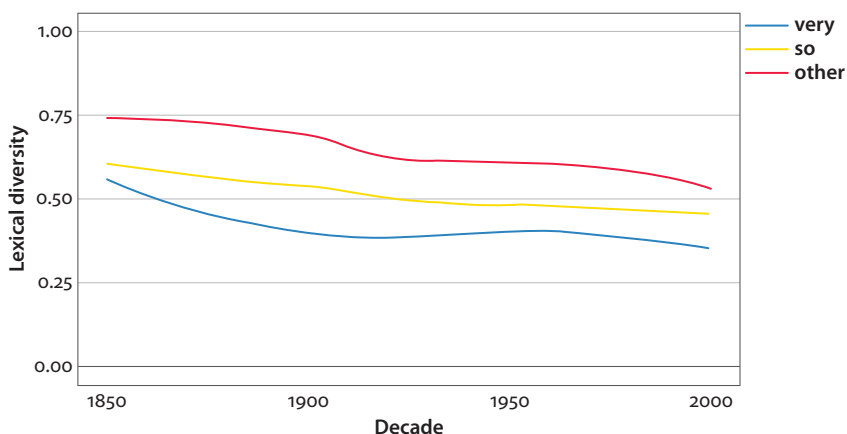


Figure 6. Lexical diversity of *very*, *so*, and all other amplifiers combined (*other*) in the fiction genre of the COHA across real time

There are at least two issues related to the current measurement and display of lexical diversity. Firstly, according to Zipf's Law, it is to be expected that the most frequent form will exhibit the lowest lexical diversity. Accordingly, a distribution in which the dominant variant has the lowest lexical diversity as operationalized here would be expected even by chance alone. However, this would also predict that the lexical diversity measure of *so* would dip and show the lowest values once it has become the most frequent form in the 1990s. However, this is not the case.

Secondly, the distributions could be affected by changes in the token frequency of adjectives. In other words, the changes in amplifier use could be confounded by changes in the frequency of adjectives. Indeed, a variant such as *so* could become dominant, not because it undergoes some kind of change, but merely because the adjectives it collocates with become more frequent. To control for changes in adjective frequency as a confounding factor, the percentages of adjectives across real time are displayed in Figure 7.

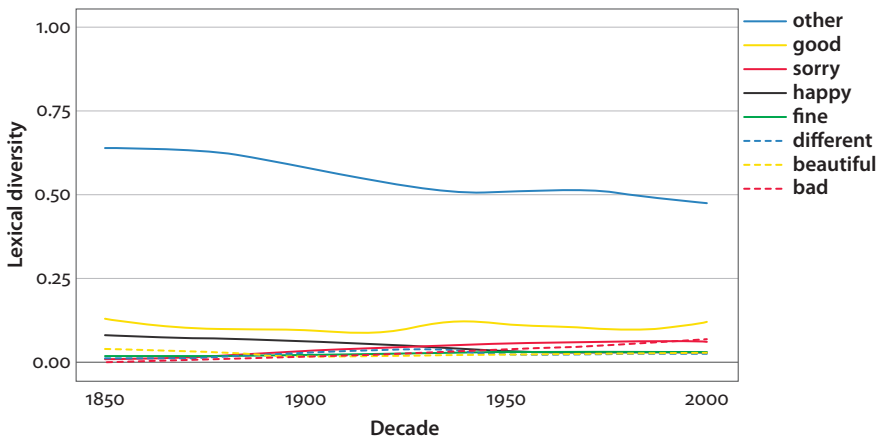


Figure 7. Distribution of adjectives across real time based on the fiction genre in COHA

Figure 7 shows that the adjective system is remarkably stable. Also note that apart from the five most frequent adjectives (*good*, *sorry*, *happy*, *fine*, and *different*), all other adjective types have been collapsed (*other*). The only trend that emerges from Figure 7 is a slow and steady decline in the relative frequency of such other adjectives. The stability of the distribution across real time indicates that changes in the frequencies of individual adjective types can be ruled out as causes for the consistency in the patterning of lexical diversity values shown in Figure 6.

We now turn to an analysis of amplifier-adjective bigrams and investigate significant collocations to gain a better understanding of the interdependency between amplifiers and adjectives. Table 2 provides an overview of significant increases and decreases of amplifier-adjective bigrams across selected decades. Table 2 also provides the p values of linear regressions that have been fit to the percentage distributions of the respective amplifier-adjective bigrams across real time.

Table 2. Significant changes in frequency of amplifier-adjective bigrams in the COHA (p-values with Simple Linear Model)

Amplifier	Adjective	1850	1880	1910	1940	1970	2000	p-value
<i>really</i>	<i>bad</i>	0.0	0.0	0.0	0.0	7.7	11.4	0.0343*
<i>really</i>	<i>good</i>	0.0	0.0	0.0	0.0	6.2	9.6	0.0347*
<i>other</i>	<i>happy</i>	6.7	7.7	15.0	29.0	20.0	35.7	0.0141*
<i>pretty</i>	<i>other</i>	0.7	0.8	3.8	4.8	4.3	5.8	0.0077**
<i>really</i>	<i>other</i>	0.7	0.8	0.8	1.8	3.4	4.1	0.0068**
<i>very</i>	<i>other</i>	55.9	54.7	55.6	54.2	42.3	30.4	0.0307*
<i>other</i>	<i>tired</i>	0.0	0.0	11.1	10.0	12.5	37.5	0.0237*

According to the linear regression analyses, only very few amplifier-adjective bigrams have increased or decreased in frequency significantly across real time. Although not as frequent as in spoken dialogue, amplifying *really* is involved in collocations that have significantly increased in frequency over time. The linear models report that the combination of *really* with *bad*, as well as with *good*, and its co-occurrence with adjectives other than *bad*, *good*, *happy*, *tired*, *beautiful*, *nice*, and *sorry* has significantly increased in frequency between 1850 and 2000. In contrast, the frequency of *very* has significantly decreased in use with adjectives other than *bad*, *good*, *happy*, *tired*, *beautiful*, *nice*, and *sorry* (the top adjectives in the current data). The adjective *happy* has increasingly been modified by amplifiers other than *pretty*, *really*, *so*, or *very* across real time. The same holds true for *tired*. This indicates that *pretty* and *really* have significantly broadened and, at least to a certain extent, compensated the decline of *very*.

To delve deeper into changes in collocational preferences, covarying collexeme analyses (Gries & Stefanowitsch 2004; Stefanowitsch & Gries 2003, 2005; see also Hilpert 2006) are applied to the data. Covarying collexeme analyses test whether elements that occur in two distinct slots within a specified construction attract or repel each other. In the present case, the first slot represents the amplifier slot and the second slot represents the adjective slot. Each slot can be occupied by a variant from a set of potential candidates – the set of amplifier variants for the first slot and the set of adjectives for the second slot. Forms attract each other if they co-occur more frequently together than would be expected by chance while they repel each other if they occur less frequently together than expected by chance. From a more technical perspective, covarying collexeme analyses provide information about whether the likelihood of an element in the first slot increases or decreases if another element occurs in the second slot. In other words, it is more likely that *nice* occurs in the second slot, given that *really* occurs in the first slot, compared with another amplifier taking the first slot. If the likelihood increases, then the forms attract each other. In the current analysis, a modified³ collexeme analysis has been implemented. The effect size measure reported here is the logged p-value (Stefanowitsch & Gries 2005). Attraction is indicated by positive values, while negative values indicate that the amplifier and the adjective repel each other (occur less frequently together than would be expected by chance).

3. The modification relates to corrections of the p-values to control the inflation of α -error rates because the analyses required to run a huge number of tests which would have resulted in false positive results (α -error) if no correction had been applied.

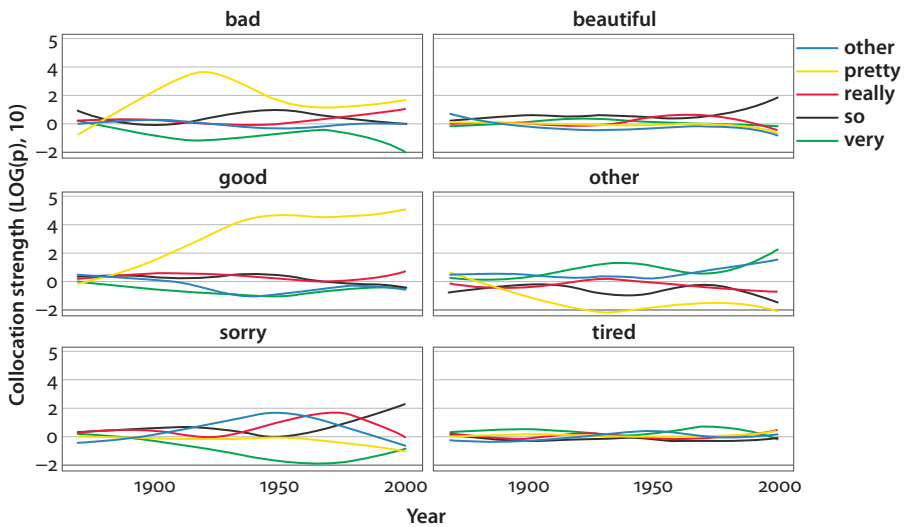


Figure 8. Collocation strength of amplifier types by adjective type across real time based on the fiction genre in COHA

The panels in Figure 8 reveal that *very* is the most general amplifier as it is preferred by adjectives other than *bad*, *beautiful*, *good*, *sorry*, and *tired* because the green line in the center right panel is consistently above 0. Amplifying *pretty* exhibits the opposite pattern as the yellow line in the center right panel is consistently below 0 – thus indicating that *pretty* is repelled by adjectives other than *bad*, *beautiful*, *good*, *sorry*, and *tired*. Indeed, *pretty* appears to have specialized on modifying with *good* as indicated by high yellow line in the center left panel. Between 1880 and 1930, *pretty* was also strongly collocating with *bad*. This collocation has, however, weakened substantially although *pretty* and *bad* still co-occur more frequently together than would be expected by chance as indicated by the wave-like patterning of the yellow line in the upper left panel. *So* has experienced an increase in collocational strength with *beautiful* and *sorry* as indicated by the upwards trend in the black line in the upper right and lower left panels. In addition, *so* appears to increasingly specialize as can be derived from the recent downwards trend of the black line in the center right panel. Given the results of the covarying collexeme analyses, the attraction between *so* and *sorry* as well as *beautiful* is quite likely the reason why *so* has replaced *very* as the dominant amplifier in the most recent data. Nonetheless, the continued specialization of *so* and its inability to occur in attributive contexts indicate that *so* will not fully replace *very*.

5. Discussion and outlook

The current analysis of change in the amplification of adjectives in predicative contexts in American English fiction data taken from the COHA has unearthed intriguing interdependencies between amplifier use and adjective frequency. The analysis of this rather substantial dataset of historical fiction data shows a steady decline in use of *very*, relative stability in the use of amplifier types in attributive contexts, and a replacement of *very* by *so* as the dominant amplifier in predicative contexts in the 1990s (see Figure 2). The results also highlight that in order to understand changes in amplifier use, one cannot consider amplifiers in isolation, but one has to take adjectives and their collocational preferences into account (see also Tagliamonte 2011: 144; Wagner 2017).

The decline of *very* in the present study mirrors trends that have been ascertained in spoken data from various other varieties such as Toronto English (Tagliamonte 2008; Tagliamonte & Denis 2014), South Eastern Ontario English (Tagliamonte & Denis 2014), North East British English (Barnfield & Buchstaller 2010; Ito & Tagliamonte 2003), and New Zealand English (D'Arcy 2015). What sets the results of the current analysis apart from previous studies is, firstly, that *so*, rather than *really*, replaces *very* in the fiction section of the COHA. Secondly, attributive contexts remain stable as a single rival variant fit to take *very*'s place is lacking in this syntactic function. Thirdly, the present analysis relies on fiction data rather than dialogue data which represents natural speech. While fiction data is more spoken- or speech-like compared with more formal written text types, it remains a written rather than a spoken genre.

The results of the second vector space model suggest that semantic similarity represents a necessary, but not a sufficient condition, for lexical replacement (see Figure 5). To elaborate, the second vector space model, as well as the strong correlation between *very* and *so*, substantiate the hypothesis that an innovative variant can only become dominant if its collocational profile is sufficiently similar to the traditional variant. However, the covarying collexeme analyses have shown that despite their general semantic similarity, *so* and *very* still exhibit distinct semantic profiles with respect to their semantic specialization. Thus, it would be inaccurate to speak of full lexical replacement which would require not only a replacement in terms of frequency but also with respect to function (D'Arcy 2015). In this respect, the semantic dissimilarity of *so* and *very* detected in the present study substantiates previous research which has argued that the successful replacement of a traditional variant by a more expressive or innovative variant is conditioned by their relative semantic similarity (see Schweinberger 2017a, 2017b). For instance, in (NZE) amplifying *really* has replaced *very* in terms of frequency (see D'Arcy 2015; Schweinberger 2017a, 2017b) while the analogous trend appears to have halted in

Irish English (IrE) (see Schweinberger 2017a, 2017b). An analysis of correlation matrices of amplifier and adjectives types in NZE and IrE has shown that *really* and *very* correlate more strongly in NZE compared with IrE. This has led Schweinberger (2017b) to hypothesize that replacement requires that the replaced and the replacing variant must be sufficiently similar for the replacement process to succeed.

Another particularly puzzling result of the present study relates to the observation that the increase in *so* in the present data is not accompanied by a broadening of its collocational profile. The lexical diversity values for *so* remain consistently above the values of *very* and do not show a significant change when *so* increased in frequency. In fact, the difference in lexical diversity appears to remain remarkably unaffected by the increase in use of *so* during or after the 1990s (see Figure 6) when *so* replaced *very* as the dominant amplifier. Linear regression models were applied to the data to test whether this finding is confounded by frequency fluctuations within the adjective system. The regression models show that the adjectival system remains remarkably stable (see Figure 7), so that changes in the frequency of certain adjective types can be ruled out as an explanation for the patterning of lexical diversity values. This consistency in the patterning of lexical diversity values which shows a stolid broadening for *so* contrasts not only with previous empirical research (e.g. Ito & Tagliamonte 2003), but also with theoretical models such as the delayed increase of discourse frequency hypothesis (DIDFH; see Mair 2004).

According to Mair's (2004) DIDFH, grammaticalization, and thus broadening of admissible contexts, precedes increases in frequency. While the replacement in the current study is arguably not a case of grammaticalization proper, the DIDFH could have explained the changes in the amplifier use. Applied to amplifiers, the DIDFH would predict that an amplifier contextually broadens by collocating with an adjectival head that it previously did not collocate with. The collocational extension of an amplifier could then be deemed analogous to a form entering new syntactic contexts in cases of grammaticalization.

Observations reported by Bybee, Perkins and Pagliuca (1994: 8) support Mair's proposed mechanism, because the findings described in Bybee et al. (ibid.) suggest that the frequency of innovative or intrusive forms increases, or even escalates, when the respective forms enter new contexts. In contrast, in the COHA data, *so* increases in use without obvious collocational or functional extension. This increase in frequency in the absence of broadening requires an explanation. One explanation which is supported by the covarying collexeme analyses is that the increase in *so* is almost exclusively dependent on its attraction to *sorry* and *beautiful*. Being a more expressive alternative to traditional *very*, *so* increased in frequency because it expanded in use in a context in which it was already admitted. The increase in frequency of *so* can be seen as a consequence of the decrease in frequency of *very* which required compensation. Therefore, the increase in *so* was not accompanied

by a broadening of contexts because it expanded in a context that *so* had already entered. This explanation ties in with the interdependency between high-frequency adjectives and amplifiers. In predicative contexts in the fiction section of the COHA, merely two adjective types appear to bear the main responsibility for the drastic increase in usage of *so* (*sorry* and *beautiful*; see Figure 8). An underlying mechanism causing this effect could therefore be that once an innovative variant co-occurs with a high-frequency adjective, it is selected for because it possesses a higher degree of expressivity compared with more frequent rival variants. This means that associating with high-frequency adjectives may cause a boost in the frequency of usage of a variant, merely because the frequency of the amplifier piggybacks on the frequency of the adjective. However, the covarying collexeme analyses also showed that *so* is not significantly collocating with the most frequent adjective in the data, *good*. The failure of *so* to collocate with *good* is likely a result of the increasing collocation of *good* with *pretty*. In fact, *pretty* appears to have blocked the amplifier-slot of *good* since the late 1870s.

The differences in the trajectories of change in the amplifier systems in other varieties and the change described here – with *so* taking on the role that *really* has taken on in other varieties – is startling. A cause for this discrepancy which takes the lack of broadening into account would be that *so* neither has nor will fully replace *very*. If true, then this would mean that *so* is experiencing a relatively short-lived and locally restricted increase in frequency due to its co-occurrence with *sorry* and *beautiful*. In this view, the increase in *so* is type-specific and limited and can, therefore, not be deemed systematic. The lack of broadening and its failure to collocate with the most frequent adjective, *good*, can be considered to support this hypothesis that *so* is rather unlikely to dominate the American English amplifier subsystem investigated here. Thus, an increase in frequency would also merely be a necessary, rather than a sufficient, condition for systematic dominance. In addition to semantic similarity and an increase in frequency, systematic dominance would also require an increase in lexical diversity (i.e. a broadening of admissible contexts and collocates).

On a methodological note, the current analysis adds to previous studies which have shown the usefulness of applying distinctive collexeme analyses for investigations into language change (Hilpert 2006). In the present study, the distinctive collexeme analyses have substantiated the inter-dependency between the collocational preferences of adjectives and ongoing changes in the use of individual amplifiers. Specifically, the application of distinctive collexeme analyses has rendered subtle tendencies visible that would otherwise have been missed. For instance, the finding that *so* is rejected by adjectives other than *bad*, *beautiful*, *good*, *sorry*, and *tired* confirms that *so* has a comparatively restrictive collocational profile. However, the opposite holds true for *very*.

One issue which remains unanswered is whether American English will follow other varieties in the sense that *really* will “become the next fully grammaticalized adjective intensifier” (Lorenz 2002: 157). Given the stability of *really* in the present (written) data, such a development seems unlikely. Also, the stability in adjective types that *really* co-occurs with indicates that it is not yet fully delexicalized. While the restriction of the change to a very specific context, that is, predicative adjectives, is unusual and contrasts with ongoing changes in other varieties, it poses questions about similarities and differences between *so* in the fiction data in American English and *really* in spoken dialogue in other varieties. In this respect, the current analysis represents a starting point for more in-depth and comparative studies on factors that cause *so* rather than *really* to become dominant in the context described in this study. But this would either require analogous analyses of sufficiently similar, and thus comparable, data sources or further analysis of register differences in American English. Both options represent possible future directions for research, but are beyond the scope of the present study. Also, one has to bear in mind that the current analysis is based on fiction data and has zoomed in on predicative contexts. It is very likely that the distribution of amplifiers and thereby the trajectory of change is substantially different in more prototypically written-like and formal genres compared with the present fiction genre data which is likely to contain passages that aim to resemble spoken dialogue data. An analysis based on a more diverse dataset would be necessary to investigate this matter further.

Finally, given that the present study focused exclusively on a subsystem of amplification, namely the amplification of adjectives by boosters, it would be intriguing to investigate the amplification of other word classes which could also be modified by the same set of amplifiers. To elaborate, the so-called GenX *so* in constructions with future going to (see (3)) has been shown to covary with alternative variants (*certainly/definitely/really*) in television dialogue (Stange 2017). Thus, a subset of the lexical amplifier forms investigated here also patterns in the amplification of verbs in the *going-to* future.

- (3) You are *so/certainly/definitely/really* going to be in trouble.

Extending the methodological approach presented here to such instances of amplification is likely to unearth patterns of variation which have so far remained undetected.

In conclusion, the results of the current study show that observable change in amplifier use is conditioned by collocational preferences of few high-frequency adjectives. The fate of amplifiers is therefore linked with the collocational preferences of such adjectives that underhandedly affect – if not determine – changes of individual amplifiers and potentially even entire amplifier (sub)systems.

References

Corpora and software

- COHA = Davies, Mark. 2010. *The Corpus of Historical American English: 400 million words, 1810–2009*. <<http://corpus.byu.edu/coha/>> (30 November 2018).
- Corpusdata = <<https://www.corpusdata.org/>> (3 June 2019).
- Hornik, Kurt. 2016. Package “OpenNLP” – Apache OpenNLP Tools Interface. <<https://cran.r-project.org/web/packages/openNLP/openNLP.pdf>> (30 November 2018).
- R Development Core Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org> (30 November 2018).

Other references

- Aijmer, Karin. 2011. Are you totally spy? A new intensifier in present-day American English. In *Marqueurs discursifs et subjectivité*, Sylvie Hancil (ed.), 155–172. Rouen: Universités de Rouen et du Havre.
- Aijmer, Karin. 2018a. That’s well bad. Some new intensifiers in spoken in British English. In *Corpus Approaches to Contemporary British English*, Vaclav Brezina, Robbie Love & Karin Aijmer (eds), 60–95. London: Routledge. <https://doi.org/10.1215/00031283-80-3-280>
- Aijmer, Karin. 2018b. Intensification with ‘very’, ‘really’ and ‘so’ in selected varieties of English. In *Corpora and Lexis*, Sebastian Hoffmann, Andrea Sand, Sabine Arndt-Lappe & Lisa Marie Dillmann (eds.), 106–139. Leiden: Brill.
- Barnfield, Katie & Buchstaller, Isabelle. 2010. Intensifiers on Tyneside: Longitudinal developments and new trends. *English World-Wide* 31(3): 252–287. <https://doi.org/10.1075/eww.31.3.02bar>
- Bauer, Laurie & Bauer, Winifred. 2002. Adjective boosters in the English of young New Zealanders. *Journal of English Linguistics* 30(3): 244–257.
- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 2007. *Longman Grammar of Spoken and Written English*, 6th edn. Harlow: Pearson.
- Bolinger, Dwight. 1972. *Degree Words*. The Hague: Mouton.
- Borst, Eugene. 1902. *Die Gradadverbien im Englischen* [Anglistische Forschungen 10]. Heidelberg: Winter.
- Breban, Tine & Davidse, Kristin. 2016. The history of *very*: the directionality of functional shift and (inter)subjectification. *English Language and Linguistics* 20(2): 221–249. <https://doi.org/10.1017/S1360674315000428>
- Brinton, Laurel J. & Arnovik, Leslie K. 2006. *The English Language. A Linguistic History*. Oxford: OUP.
- Bybee, Joan L., Perkins, Revere D. & Pagliuca, William. 1994. *The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World*. Chicago, IL: University of Chicago Press.
- Calle-Martín, Javier. 2014. On the history of the intensifier *wonder* in English. *Australian Journal of Linguistics* 34(3): 399–419. <https://doi.org/10.1080/07268602.2014.898224>
- D’Arcy, Alexandra F. 2015. Stability, stasis and change – The *longue durée* of intensification. *Diachronica* 32(4): 449–493. <https://doi.org/10.1075/dia.32.4.01dar>
- Fettig, Adolf. 1934. *Die Gradadverbien im Mittelenglischen* [Anglistische Forschungen 79]. Heidelberg: Winter.

- Fuchs, Robert. 2016. Register variation in intensifier usage across Asian Englishes. In *Discourse-Pragmatic Variation and Change: Insights from English*, Heike Pichler (ed.), 185–213. Cambridge: CUP.
- Fuchs, Robert. 2017. Do women (still) use more intensifiers than men? Recent change in the sociolinguistics of intensifiers in British English. *International Journal of Corpus Linguistics* 22(3): 345–374. <https://doi.org/10.1075/ijcl.22.3.o3fuc>
- Gries, Stefan T. & Stefanowitsch, Anatol. 2004. Extending collostructional analysis: A corpus-based perspectives on ‘alternations’. *International Journal of Corpus Linguistics* 9(1): 97–129. <https://doi.org/10.1075/ijcl.9.1.o6gri>
- Hilpert, Martin. 2006. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory* 2(2): 243–256. <https://doi.org/10.1515/CLLT.2006.012>
- Ito, Rika & Tagliamonte, Sali. 2003. *Well weird, right dodgy, very strange, really cool*: Layering and recycling in English intensifiers. *Language in Society* 32, 257–279. <https://doi.org/10.1017/S004740450322055>
- de Klerk, Vivian. 2005. Expressing levels of intensity in Xhosa English. *English World-Wide* 26(1): 77–95. <https://doi.org/10.1075/eww.26.1.o4dek>
- König, Ekkehard. 2017. The comparative basis of intensification. In *Exploring Intensification. Synchronic, Diachronic and Cross-linguistic Perspectives* [Studies in Language Companion Series 189], Maria Napoli & Miriam Ravetto (eds), 15–32. Amsterdam: John Benjamins.
- Labov, William. 1966. *The Social Stratification of English in New York City*. Washington DC: Center for Applied Linguistics.
- Labov, William. 1985. Intensity. In *Meaning, Form and Use in Context: Linguistic Applications*, Deborah Schiffrin (ed.), 43–70. Washington, DC: Georgetown University Press.
- Levshina, Natalia. 2015. *How to do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.
- Lorenz, Gunter R. 2002. Really worthwhile or not really significant: A corpus-based approach to the delexicalisation and grammaticalisation of adverbial intensifiers in Modern English. In *New Reflections on Grammaticalization* [Typological Studies in Language 49], Ilse Wischer & Gabriele Diewald (eds), 143–161. Amsterdam: John Benjamins.
- Macaulay, Ronald. 2002. Extremely interesting, very interesting, or only quite interesting: Adverbs and social class. *Journal of Sociolinguistics* 6(3): 398–417.
- Macaulay, Ronald. 2006. Pure grammaticalization: The development of a teenage intensifier. *Language Variation and Change* 18(3): 267–283. <https://doi.org/10.1017/S0954394506060133>
- Mair, Christian. 2004. Corpus linguistics and grammaticalisation theory. Statistics, frequencies and beyond. In *Corpus Approaches to Grammaticalization in English* [Studies in Corpus Linguistics 13], Hans Lindquist & Christian Mair (eds), 121–150. Amsterdam: John Benjamins.
- Méndez-Naya, Belén. 2003. On intensifiers and grammaticalization: The case of *swipe*. *English Studies* 84(4): 372–391. <https://doi.org/10.1076/enst.84.4.372.17388>
- Méndez-Naya, Belén. 2008. Introduction. Special issue on English intensifiers. *English Language and Linguistics* 12(2): 213–219. <https://doi.org/10.1017/S136067430800259>
- Méndez-Naya, Belén & Pahta, Päivi. 2010. Intensifiers in competition: The picture from early English medical writing. In *Early Modern English Medical Texts: Corpus Description and Studies*, Irma Taavitsainen & Päivi Pahta (eds), 191–214. Amsterdam: John Benjamins.
- Mustanoja, Tauno F. 1960. *A Middle English Syntax, Part 1: Parts of Speech*. Helsinki: Société Néophilologique.
- Nevalainen, Terttu. 2008. Social variation in intensifier use: Constraint on *-ly* adverbialization in the past? *English Language and Linguistics* 12(2): 289–315. <https://doi.org/10.1017/S1360674308002633>

- Nevalainen, Terttu & Rissanen, Matti. 2002. Fairly pretty or pretty fair? On the development and grammaticalization of English downtoners. *Language Sciences* 24: 359–380.
[https://doi.org/10.1016/S0388-0001\(01\)00038-9](https://doi.org/10.1016/S0388-0001(01)00038-9)
- Núñez-Pertero, Paloma & Palacios Martínez, Ignacio. 2014. *That's absolutely crap, totally rubbish*. The use of intensifiers *absolutely* and *totally* in the spoken language of British adults and teenagers. *Functions of Language* 21(2): 210–237. <https://doi.org/10.1075/fol.21.2.03pal>
- Palacios Martínez, Ignacio & Núñez-Pertero, Paloma. 2012. He's absolutely massive. It's a super day. Madonna, she is a wicked singer. Youth language and intensification: A corpus-based study. *Text and Talk* 32(6): 773–796. <https://doi.org/10.1515/text-2012-0036>
- Paradis, Carita. 1997. *Degree Modifiers of Adjectives in Spoken British English*. Lund: Lund University Press.
- Paradis, Carita. 2008. Configurations, construals and change: Expressions of DEGREE. *English Language and Linguistics* 12(2): 317–343. <https://doi.org/10.1017/S1360674308002645>
- Partington, Alan. 1993. Corpus evidence of language change: The case of intensifiers. In *Text and Technology: In Honour of John Sinclair*, Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds), 177–192. Amsterdam: John Benjamins.
- Peters, Hans. 1992. English boosters: Some synchronic and diachronic aspects. In *Diachrony within Synchrony: Language History and Cognition*, Günter Kellermann & Michael D. Morrissey (eds), 529–545. Frankfurt: Peter Lang.
- Peters, Hans. 1993. *Die englischen Gradadverbien der Kategorie booster*. Tübingen: Narr.
- Peters, Hans. 1994. Degree adverbs in Early Modern English. In *Studies in Early Modern English*, Dieter Kastovsky (ed.), 269–288. Berlin: Mouton de Gruyter.
- Quirk, Randolph, Greenbaum, Sydney, Leech, Geoffrey & Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rickford, John, Wasow, Thomas, Zwicky, Arnold & Buchstaller, Isabelle. 2007. Intensive and quotative all: Something old, something new. *American Speech* 82(1): 3–31.
<https://doi.org/10.1215/00031283-2007-001>
- Rissanen, Matti. 2008. From 'quickly' to 'fairly': On the history of *rather*. *English Language and Linguistics* 12(2): 345–359. <https://doi.org/10.1017/S1360674308002657>
- Schweinberger, Martin. 2017a. *Using intensifier-adjective collocations to determine mechanisms of change*. Paper presented at ICAME 38 (International Computer Archive of Modern and Medieval English). Prague, Czech Republic, 24–28.05.2017. <<http://www.martinschweinberger.de/docs/ppt/schweinberger-ppt-prague-2017-05-27.pdf>> (30 November 2018).
- Schweinberger, Martin. 2017b. *Modelling intensification in New Zealand English data*. Paper presented at Methods in Dialectology XVI. Tokyo, Japan, 7–11.08.2017. <<http://www.martinschweinberger.de/docs/ppt/schweinberger-ppt-tokyo-2017-08-10.pdf>> (November 30 2018).
- Stange, Ulrike. 2017. "You're so not going to believe this": The use of GenX *so* in constructions with future *going to* in American English. *American Speech* 92(4): 487–524.
<https://doi.org/10.1215/00031283-4395168>
- Stefanowitsch, Anatol. 2010. Empirical cognitive semantics: Some thoughts. In *Quantitative Cognitive Semantics: Corpus-driven Approaches*, Dylan Glynn & Kerstin Fischer (eds), 355–380. Berlin: Mouton de Gruyter.
- Stefanowitsch, Anatol & Gries, Stefan T. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.
<https://doi.org/10.1075/ijcl.8.2.03ste>
- Stefanowitsch, Anatol & Gries, Stefan T. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1): 1–43. <https://doi.org/10.1515/cllt.2005.1.1.1>

- Stenström, Anna-Brita. 1999. He was really gormless – she's bloody crap: Girls, boys and intensifiers. In *Out of Corpora: Studies in Honour of Stig Johansson*, Hilde Hasselgård & Signe Oksefjell (eds), 69–78. Amsterdam: Rodopi.
- Stoffel, Cornelis. 1901. *Intensives and Downtoners* [Anglistische Forschungen 1]. Heidelberg: Winter.
- Tagliamonte, Sali A. 2008. So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language & Linguistics* 12(2): 361–394. <https://doi.org/10.1017/S1360674308002669>
- Tagliamonte, Sali A. 2011. *Variationist Sociolinguistics. Change, Observation, Interpretation*. Malden, MA: Wiley-Blackwell.
- Tagliamonte, Sali A. & D'Arcy, Alexandra. 2009. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language* 85(1): 58–108.
- Tagliamonte, Sali A. & Denis, Derek. 2014. Expanding the transmission/diffusion dichotomy: Evidence from Canada. *Language* 90(1): 90–136. <https://doi.org/10.1353/lan.2014.0016>
- Tagliamonte, Sali & Roberts, Chris. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series *Friends*. *American Speech* 80(3): 280–300. <https://doi.org/10.1215/00031283-80-3-280>
- Tao, Hongyin. 2007. A corpus-based investigation of absolutely and related phenomena in spoken American English. *Journal of English Linguistics* 35(1): 5–29. <https://doi.org/10.1177/0075424206296615>
- Wagner, Susanne. 2017. Totally new and pretty awesome: Amplifier-adjective bigrams in GloWbE. *Lingua* 200: 63–83. <https://doi.org/10.1016/j.lingua.2017.08.004>
- Xiao, Richard & Tao, Hongyin. 2007. A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic Studies* 1(2): 241–273. <https://doi.org/10.1558/sols.vi12.241>

The development and pragmatic function of a non-inference marker

That is not to say (that)

Laurel J. Brinton

University of British Columbia, Vancouver

That is not to say (that) is an introductory clause refuting an inference that could be drawn from the previous discourse; it often occurs in a negative context (with a positive reading) and is characteristic of written genres, with lowest frequency in fiction. The earliest examples date from the 16th century but the next examples found date from the mid- to late-19th century. The gap in the corpus data may be explained by the predominantly fiction/drama make-up of available corpora. The development of *that is not to say (that)* is not an entirely prototypical case of grammaticalization as certain parameters (loss of *that*, contracted forms) are inconclusive and there are no 'lexical' uses of the form and hence no 'divergence'.

1. Introduction

The clause *that is (to say)* functions as a marker of apposition, introducing explanation, elaboration, particularization, or exemplification. It is treated in standard grammars (e.g. Biber et al. 1999: 876–877; Huddleston & Pullum et al. 2002: 779; Quirk et al. 1985: 637–638, 1307–1308, etc.) and defined in the *Oxford English Dictionary* (OED, s.v. *say*, v.¹ and int., def. P2a (a); s.v. *that*, pron.¹, adj., and adv., def. A I 5a). *That is* originates in Old English, possibly as a calque on Latin *id est*; the longer form *that is to say* appears later in Old English and becomes established by Middle English, perhaps reinforced by or calqued on French *c'est-à-dire* (see Brinton 2008: 104–109). The alternative *which is to say* also appears in Middle English, with a number of other variants found: *that/it is to mean/understand/wit, to wit* (see Koivisto-Alanko & Rissanen 2002); cf. Latin *videlicet* (Moore 2006), *scilicet*.

In contrast, the negative counterpart, *that/this is not to say (that)*, receives no treatment in standard grammars and is omitted from the OED (though the quotation database contains 27 examples). The construction is mentioned in passing by Goldberg and van der Auwera (2012) as a fixed phrase, with the observation that

this is neither the subject nor the object of *to say*. Online discussions point out that the construction is used “to make sure the person you are talking to does not think something that is not true” (*Longman Dictionary*) or “for adding a statement that corrects what you have just said” (*Macmillan Dictionary*) and may be defined as “that does not mean” (*Merriam-Webster Dictionary*).

This chapter provides a corpus-based pragmatic analysis of *that is not to say* (*that*). Section 2 discusses the variant forms of the construction and their distribution in varieties and genres of English. It also analyzes the synchronic status of *that is not to say* as a discourse-pragmatic element. Section 3, after establishing the dating of the construction in historical corpora, discusses the degree of grammaticalization exhibited by this construction. Gaps in the corpus data are also explained. Section 4 concludes the chapter.

2. Form and function of *that is not to say* in Present-day English

That is not to say (*that*) exists in a number of variant forms, with *that*, *this*, *which*, and *it* as subject, with either verb contraction or negative contraction, and with or without complementizer *that*, as shown by the examples from the *Corpus of Contemporary American English* (COCA) in (1). These variations set the negative construction apart from its positive counterpart, where the subject demonstrative *that* is almost universal, contracted *that’s to say* is rare, and complementizer *that* is almost never present.

- (1) a. *That is not to say that* we don’t have discussions and conversations.
(COCA: 2012 Fox_Hannity)
- b. *that’s not to say* it could not produce criminal charges.
(COCA: 2017 PBS_Newshour)
- c. *This isn’t to say* there haven’t been a few heartening successes in recent years.
(COCA: 2008 Newsweek)
- d. *This is not to say that* no consensus can be achieved in these areas today.
(COCA: 2017 *The American Bankruptcy* [sic] *Law Journal*)
- e. I haven’t gotten any blowback for it yet, *which is not to say* it won’t happen.
(COCA: 2017 *The Verge*)
- f. but they never believed in it. *Which isn’t to say* they didn’t want it.
(COCA: 1997 Newsweek)
- g. *It’s not to say* it would have stopped everything, but it would have stopped some.
(COCA: 2012 *New York Times*)
- h. *It is not to say that* this theory is essentially atheistic or represents atheism.
(COCA: 1991 *Journal of Church and State*)
- i. *This was not to say* that Fisher wasn’t good-looking or attractive.
(COCA: 2012 DeFelice, *The Helios Conspiracy*)

For convenience I will refer to all variants of the present tense as the *that is not to say* construction. Rare past tense forms (as in (1i)) will not be considered in this chapter.

Corpus evidence points to the use of *that is not to say* as an introductory clause (or in the case of the *which*-forms, a sentential relative clause) refuting an inference that could be drawn from the previous discourse and thus it may be understood as a non-inference marker. For example, if we look at the larger context of (1a), given in (2), we see that Michelle Obama is negating a possible inference – from “he has dozens of really smart people who surround him” – that she and Barack Obama do not converse about important issues.

- (2) M-OBAMA: I am one of his biggest confidants. But he has dozens of really smart people who surround him. *That is not to say that* we don’t have discussions and conversations. *That is not to say that* my husband doesn’t know how I feel.
(COCA: 2012 Fox_Hannity)

That is not to say frequently occurs in the context of a following negative, where it receives a positive reading (as in (1a)–(1f)). Liberman (2012) describes the construction as a “convoluted litotic way” of expressing the positive: “the author is taking a position that might seem to oppose X, and wants to explicitly endorse X in order to prevent misunderstanding”. In fact, *that is not to say* is typically accompanied by a denial of the possible inference, sometimes marked explicitly (with *on the contrary*, *rather*, *but*, *at best*, and so on) as in (3).¹

- (3) a. *This is not to say* that these cases are easy; *on the contrary*, many are quite complex.
(COCA: 2016 Iowa Law Review)
b. *This is not to say* that the urban peripheries are homogeneously poor. *Rather*, the slums are socio-economically heterogeneous
(COCA: 2008 Anthropological Quarterly)
c. *That is not to say* that Ted Hughes drove her to suicide, *but* it seems to me that it was a tragic match,
(COCA: 1998 New York Times)
d. it begins to speak, *which is not to say* to communicate, *but rather* to iterate.
(COCA: 2006 Art Bulletin)
e. *That’s not to say* that there is any kind of consensus in Catalonia. *At best*, a slim majority backs independence,
(COCA: 2017 Slate Magazine)

Liberman (2012) has discussed the existence of *triple negatives* in the *that is not to say* construction, as in (4).

- (4) *That is not to say* that I don’t think that corpus work can’t be useful, even in theoretical syntax.
(Briscoe 2000, Interview of Gazdar)

1. As noted by a reviewer, none of these show the presence of a negative in the complement clause, most likely because the negative is an inherent part of the markers *on the contrary*, *rather*, etc.

He finds other (rare) internet examples of triple negatives, all involving *that is not to say* + *I don't say* + negative complement clause. He notes that, logically, an example such as this should be interpreted as negative; thus, Example (4) should be read 'corpus work can't be useful'. However, it is clearly intended as, and interpreted as, positive ('corpus work can be useful'). Unable to explain this interpretation, Liberman sees this as a form of *misnegation*. He suggests four possible answers:

- we are unable to deal with complex combinations of certain logical operators;
- the connection between English and modal logic may involve unexpected ambiguities;
- negative concord is alive and well in English; or
- odd things become idioms or verbal habits.

While Liberman does not resolve the problems raised here (nor do commentators on the blog), he tends toward an analysis involving negative concord.

Such triple negatives are extremely rare and perhaps seen as mistakes, though interestingly, there appears to be a difference in interpretation, depending on the syntax of the third negative. Thus, Examples (5a)–(b), where the third negative occurs in a complement clause, as in Example (4), are read as positive (e.g., 'the children had been worth it', 'a stay at home mother is a job'), whereas Examples (5c)–(f), where the third negative occurs in an adjectival (relative) clause, are read as negative (e.g., 'there were some men who did not embrace the concept', 'we did have some who were not on board').

- (5) a. *It's not that* he didn't think that the children hadn't been worth it – all his children. (COCA: 2017 Fresh Air)
- b. *This is not to say* I don't think being a stay at home mother is *not* a job. (2014; www.mtv.com/news/1978048/amy-poehler-yes-please-book-feminist-quotes/)
- c. *This is not to say* there weren't some men in the department who did *not*, shall we say, embrace the concept of having women out on patrol. (COCA: 1992 CNN_Sonya)
- d. ... *which is not to say* that there are *not* members of this Cabinet who are *not* to the left. (COCA: 1992 ABC_Brinkley)
- e. *That's not to say* we didn't have some who were *not* totally on board. (COCA: 2014 Christianity Today)
- f. *Which is not to say* there aren't sophisticated people out there who aren't "curious" about people who have done things such as I. (2011 Belber, *Dusk rings the bell*; <<https://books.google.ca/books?isbn=0822224925>>)
- g. *This is not to say* that there aren't times when fires don't have a mind of their own. <<https://www.fs.usda.gov/main/hmnf/learning/safety-ethics>>

It should be noted that in the *that is not to say* construction, *say* has a weakened sense, not ‘to express in words’ but ‘to convey information without using words; to indicate’ (OED, s.v. *say*, v.¹ and int., def. III). A reduced variant of the *that is not to say* construction without “to say” can also be found, as shown in (6).

- (6) a. *It is not that* there are not interesting underlying changes in the categories,
(COCA: 1997 *Canada-United States Law Journal*)
- b. *It is not that* he finds expensive wines to be wrong.
(COCA: 2009 *San Francisco Chronicle*)
- c. *It's not that* they are bad people; it's that they feel entitled.
(COCA: 2017 *MarketWatch*)
- d. *It isn't that* government is all bad, as Carson stated, but that government is doing things it should not be doing.
(COCA: 2017 *Omaha World-Herald*)
- e. Admit guilt. How interesting. *It's not that* I haven't thought about that
(COCA: 2017 CBS: 48 Hours)
- f. *It is not that* she has given up on her mission, but rather, she has gained a sense of the island's diversity and rhythms
(COCA: 2010 *Journal of Film & Video*)

While this construction often marks non-inference, in many instances this meaning cannot be clearly distinguished from the more literal meaning, ‘it is not the case that’. The non-inference meaning seems to be most natural in cases where the *that*-clause expresses a subjective opinion (as in (6a)–(d)) and less likely in cases where it describes a state of affairs (as in (6e)–(f)). Unlike *that is not to say*, complementizer *that* is obligatorily present in this form: **It is not he finds expensive wines to be wrong*.

Because of the difficulties of interpretation, this construction is not included in the quantitative data given in Sections 2.1 and 2.2.2 and only briefly considered in the historical development of the construction (see Section 3.1).

2.1 Distribution of *that is not to say*

That is not to say is used with comparable frequency in British and American English. In COCA, the overall frequency is 4.19/million, while in the *British National Corpus* (BYU-BNC) it is 4.46/million. As shown in Figure 1, where all variants of particular types are collated, we see that there would appear to be a preference for *this* over *that* in British English (56% to 35%) and a smaller preference for *that* over *this* in American English (44% to 38%). Both *which is not to say* and *it is not to say* combined constitute under 20% of the variants in both varieties.²

2. Examination of the 1990–1994 data from COCA (closer in date to the BNC data) reveals a more comparable overall frequency of 4.62/million, and likewise a preference for *this* over *that*

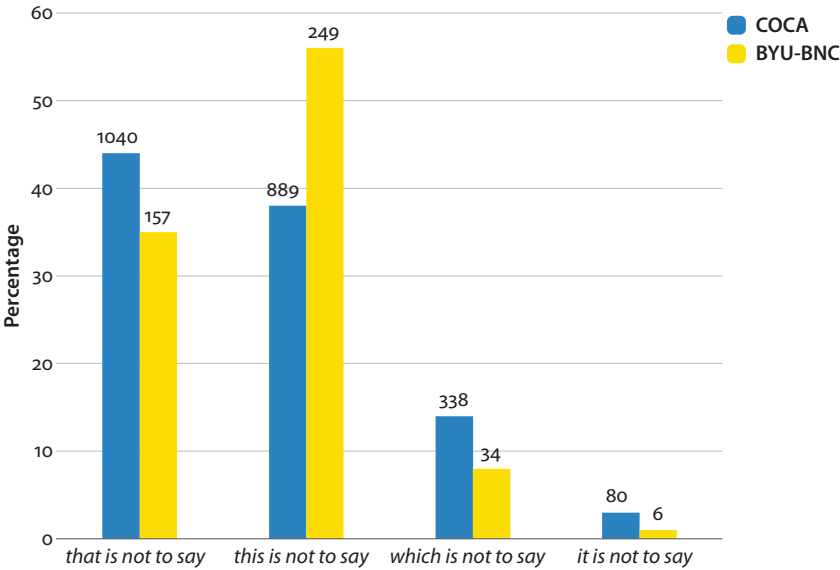


Figure 1. Forms of the *that is not to say* construction in COCA and BYU-BNC, percentages and raw frequencies

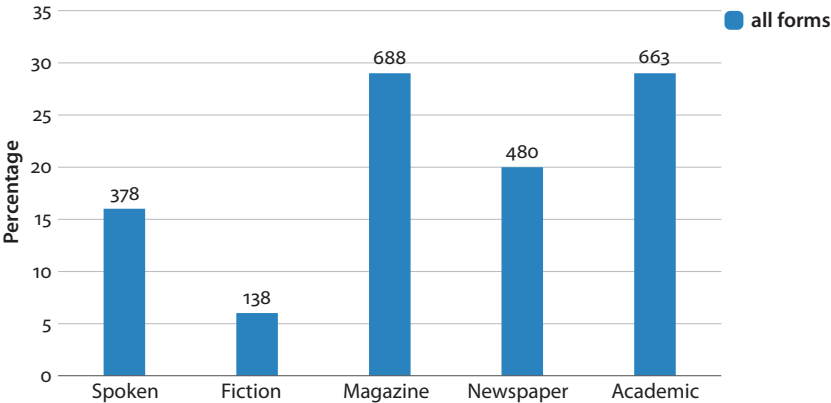


Figure 2. Distribution of all variants of the *that is not to say* construction by genre in COCA

(48% to 37%). While no clear trends can be concluded from the short time span of COCA, it would appear that there is a decline in overall frequency of the construction in American English, with increasing preference for *that*.

The distribution of *that is not to say* by genre in COCA (see Figure 2) reveals that the usage rates of *that is not to say* in Fiction are very low (6%) and those in Spoken are also quite low (16%). The construction is clearly associated with the written or text-based genres, Magazine (29%), Newspaper (20%), and Academic (29%). A similarly low use of the construction in Fiction (3%) and Spoken (6%) is evident in the BYU-BNC as well. The scarcity of *that is not to say* in Fiction poses difficulties for diachronic study, as will be discussed below (see Section 3.5).

2.2 Synchronic status

The status of *that is not to say* as a matrix clause – or not – can be approached from the perspective of either function or form. Formally, the presence or absence of *that* is considered critical.

2.2.1 Discourse-pragmatic function

Assuming different perspectives and focusing on a diverse set of forms, a number of scholars have argued that certain syntactic matrix clauses should be interpreted as functioning on the discourse-pragmatic plane in a secondary function. I suggest that these analyses can be extrapolated to the *that is not to say* construction.

In a corpus of conversational English, Thompson (2002) identifies what she calls CTPs, or *complement-taking predicates*, including first-person subjects with affirmative and negative verbs (*I think, I know, I see, I guess, I remember*) as well as forms such as *it doesn't matter, it's too bad, it's interesting, it's a shame* (2002: 138). She argues that a CTP functions as a “reusable fragment” (2002: 141) “expressing epistemic, evidential, or evaluative stance” or “negotiating alignment” (2002: 131); it serves as a “frame” which introduces or “projects” the stance of the following utterance (cf. Hopper & Thompson 2008). The declarative clause to which it is linked may or may not be preceded by *that*, but importantly does not serve as the “complement” to the CTP and is not “subordinate”. CTPs “do not constitute the speakers’ interactional agenda”; their “profile” is overridden by the “apparent” complement clause (Thompson 2002: 134). Typically, CTPs may function as parentheticals in initial, medial, or final position or stand alone and have the status of a “discourse marker *par excellence*” (2002: 144). Although it does not meet the criterion of movability, *that is not to say* would seem to belong to the Thompson’s class of CTPs.

Discussing a range of constructions posing difficulties for the segmentation of discourse, Verhagen (2001; see also Schilperood & Verhagen 1998) argues that Dutch *daarbij komt (dat)* ‘to this should be added (that)’ is not conceptually complete, nor even interpretable, without the information provided in the subordinate clause. *Daarbij komt* provides specifications on what Verhagen calls the

coordination dimension (i.e. intersubjective coordination between speaker and hearer), explicitly marking the source of subjectivity. In contrast, the subordinate clause supplies information on the content dimension and is necessary to make the matrix clause conceptually complete. *Daarbij komt* thus does not count as an independent discourse segment on the content dimension. As Verhagen notes, this analysis turns the “traditional notion of ‘dependent clause’ upside down, by showing that it is the matrix clause that is actually conceptually dependent on a subordinate one” (2001: 349). He points out further that constructions such as these are widespread, including a range of complement-taking constructions (2001: 345). The *that is not to say* construction bears close similarity to Verhagen’s class of conceptually dependent matrix clauses.

Boye and Harder (2012) likewise recognize two levels: structure and usage (discourse). Within this dual structure, forms are either grammatical or lexical (nonaddressable or addressable), and discursively primary or secondary (discourse central/foregrounded/main point or ancillary/backgrounded). Using this classification, we can see *that is not to say* as grammatical and secondary. First, it is grammatical rather than lexical because it does not contribute to “the primary element of utterance meaning” (Boye & Harder 2012: 11). It is nonfocalizable and nonaddressable (2012: 14–16). Among other things, it cannot occur in focus position in a cleft or pseudo-cleft sentence and it cannot be questioned by a *wh*-question.³ Second, *that is not to say* does not convey prominent mental content and is hence secondary in relation to other expressions (2012: 7–8).

Discussing CTPs, Boye and Harder (2007) note that when grammatical, they are non-argument-assigning and not-predicating. They suggest a number of features of (weakly) grammaticalized CTPs (2007: 578–580). *That is not to say* meets four of the five criteria proposed by Boye and Harder (2007):⁴

1. Only a grammatical CTP allows NEG-raising (i.e. narrow scope):⁵ for instance, *That is not to say he’s happy* is understood as raising from *That is to say he’s not happy*.

3. For example, a grammatical element such as *I think* in *He is rich, I think* cannot be in focal position in a cleft (**It is that he is rich that I think*) and cannot be questioned by *Really?* (this can only question whether he is rich or not). *That is not to say he’s rich* would exhibit the same restrictions.

4. Although Boye and Harder (2007) define the CTP narrowly in English (only weak assertive predicates such as *think, believe, suppose, expect* in the present tense, indicative, declarative), they admit that in Danish the category may include, among others, utterance verbs (*say*) or perception verbs (*hear*).

5. The grammatical *I don’t think I love her* involves raising from *I think I don’t love her*, but the lexical *I don’t regret that he’s rich* is not the raised version of *I regret that he’s not rich*.

2. In a grammatical CTP, a tag refers to subject of complement clause not to the subject of the main clause: *That is not to say that he's happy, is he?* **It's not to say he's happy, is it?*⁶
3. A grammatical CTP does not allow adverbial modification: **That is not to say conclusively he's not happy.*⁷
4. It allows a limited range of morphological variants (e.g. a low number of past tense forms) and syntactic combinations; for instance, it cannot combine with (indirect) imperatives or interrogatives: **That is not to say, go home.* **That is not to say, is he happy?*

That is not to say does not meet the criterion of having adverbial-like distribution, that is, of appearing in medial or final position. It is important to note that Boye and Harder do not see these five as defining criteria but rather as symptoms of grammaticalization, where grammaticalization is a matter of degree: “we do not exclude the existence of grammaticalized CTP variants that fail to show one or more of these symptoms” (2007: 583).

One puzzle to be discussed below in respect to the historical development of *that is to say* is the lack of a purely lexical variant. That is, the construction can never be addressable, nor can it be the discourse central or foregrounded element: *That is not to say he's happy.* **Isn't it?*

In conclusion, these arguments would point to the status of *that is not to say* as dependent or subordinate on the discourse-pragmatic level – as a frame or discourse marker in Thompson's (2002) view, as conceptually dependent in Verhagen's (2001) schema, and as discursively secondary in Boye and Harder's (2007, 2012) terms.

2.2.2 Presence or absence of *that*

In the *that is not to say* construction, *that* is present in the majority of instances, although there is some variability (see Tables 1 and 2). The only cases (apart from some forms with very low frequency) in which *that* is more often absent are the contracted forms *that's not to say* (in COCA and BYU-BNC) and *which isn't to say* (in COCA).

6. Tag questions in this context would appear to be very rare (no examples were found in COCA); thus, grammaticality judgments remain somewhat tenuous.

7. I found one example in COCA: *This is not to say cynically that all party principles are disposable* (COCA: 1992 *American Heritage*). The restriction of adverbial modification is not absolute, even in the most highly grammaticalized CTPs. For example, sentence-final *I think* permits adverbial modification in a handful of cases in COCA: *There's a lot of lazy people on board, I just think.* (COCA: 2002 *Atlantic Journal Constitution*).

Table 1. Complementizer *that* in the *that is not to say* construction in COCA

	Ø S		<i>that</i> S		Total
<i>that is not to say</i>	86	(32%)	183	(68%)	269
<i>that isn't to say</i>	18	(47%)	20	(53%)	38
<i>that's not to say</i>	445	(61%)	286	(39%)	731
<i>that ain't to say</i>	2	(100%)	–		2
<i>this is not to say</i>	142	(17%)	673	(83%)	815
<i>this isn't to say</i>	24	(32%)	50	(68%)	74
<i>which is not to say</i>	116	(44%)	149	(56%)	265
<i>which isn't to say</i>	45	(63%)	26	(37%)	71
<i>which ain't to say</i>	2	(100%)	–		2
<i>it is not to say</i>	4	(57%)	3	(43%)	7
<i>it isn't to say</i>	–		2	(100%)	2
<i>it's not to say</i>	31	(44%)	40	(56%)	71
Total	915	(39%)	1,432	(61%)	2,347 (100%)

Table 2. Complementizer *that* in the *that is not to say* construction in BYU-BNC

	Ø S		<i>that</i> S		Total
<i>that is not to say</i>	28	(29%)	70	(71%)	98
<i>that isn't to say</i>	4	(44%)	5	(56%)	9
<i>that's not to say</i>	32	(64%)	18	(36%)	50
<i>this is not to say</i>	24	(10%)	223	(90%)	247
<i>this isn't to say</i>	–		2	(100%)	2
<i>which is not to say</i>	8	(28%)	21	(72%)	29
<i>which isn't to say</i>	2	(40%)	3	(60%)	5
<i>it is not to say</i>	–		1	(100%)	1
<i>it isn't to say</i>	1	(100%)	–		1
<i>it's not to say</i>	2	(50%)	2	(50%)	4
Total	101	(23%)	345	(77%)	446 (100%)

The presence of *that* may reflect on the status of the *that is not to say* construction and, as we will discuss later, its stage of development (see Section 3.3). In the literature on epistemic parentheticals such as *I think* and *I guess*, *that* has traditionally been understood as pointing to the matrix clause status of these forms. Thus, Thompson and Mulac (1991: 313) argue that *I think* in (7a) is a matrix clause, in (7b) it is indeterminate between matrix and parenthetical (and the bridging context for grammaticalization), and in (7c) it is fully parenthetical.

- (7) a. *I think* that we're definitely moving towards being more technological.
- b. *I think* Ø exercise is really beneficial, to anybody.
- c. It's just your point of view you know what you like to do in your spare time
 I think.

However, increasingly it has been argued that the presence or absence of *that*⁸ is not a determining factor in the status of *I think* or *I guess* as matrix clause or comment clause (see Brinton 2008 on comment clauses). For Kearns (2007), deletion of *that* does not indicate a change in syntactic status but rather in informational prominence, where the embedded clause achieves prominence over the matrix; signs of the reduced prominence of the matrix are its omissibility, its possible replacement by an epistemic adverb, its occurrence in medial or final position, and its lack of control over the form of tags. For others, however, the presence of *that* is of no consequence to either syntactic or discourse status. Aijmer (1997) sees tag behavior as the distinguishing criterion for comment clause status, not the presence or absence of *that*. Thompson (2002: 146) notes that “even clauses marked in the ways that complements may be marked, with the linker *that*, can be used as [CTPs]”. Torres Cacoullós and Walker (2009) argue that the autonomy of even the most highly grammaticalized discourse formulas is incomplete; they may retain the use of *that* as a trace of their source construction (i.e. Hopper’s 1991 principle of persistence). For Dehé and Wichmann (2010), prosody is the determinant, not the presence of *that*: whether prosodically integrated or not, matrix-clause uses carry accent on the pronoun, while comment-clause uses have accent on the verb. *That* may join the pronoun-verb sequence (suggesting a high degree of routinization with *that* not acting as a complementizer) or it may join the following domain.⁹ In Kaltenböck’s (2011) corpus, *I think that* and *I think* Ø show similar prosodic patterns, with the occurrence of *that* in fact correlating with reduced prosodic prominence and secondary function. He argues that *that* in these cases is not functioning as a marker of subordination but rather as a filler.

If we assume that the *that is not to say* construction must be interpreted as discursively secondary (see Section 2.2.1), the presence of complementizer *that* may be determined by factors other than syntactic or discourse status. For example, Kaltenböck (2011: 108) notes that complementizer *that* distances the speaker from the content of the embedded clause and is common with CTPs with more specific semantic content. Similarly, Torres Cacoullós and Walker (2009: 34) point out that *that* demarcates the clauses and is more common when there is more lexical content

8. A number of syntactic, discursive, and stylistic factors are known to favor the deletion of *that* in complementation structures, including complexity of the matrix clause, linear adjacency of the matrix verb and *that*, full NP or pronominal status of the clause subjects, and so on (see, e.g., Biber et al. 1999: 680–683; Finegan & Biber 1995; Rissanen 1991). There are also significant differences in *that*-deletion among different verbs.

9. A discourse marker for Dehé and Wichmann (2010) consists of cases of semantically bleached and grammaticalized *I believe/think* serving a discourse (mitigating, politeness) function; these are prosodically integrated and lack stress.

in the two clauses. Increased lexical or semantic content may be true of *that is not to say*, though this is hard to quantify. The lower use of *that* with the contracted forms is to be expected if we see these forms as more highly grammaticalized, as is discussed below (Section 3.4), and more informal.

3. Historical development of *that is not to say*

Because of the low frequency of the *that is not to say* construction, a large number of historical corpora were examined. Among those providing examples are the corpora listed in Table 3.

Table 3. Historical corpora used in this study

American	Years	Words
<i>The Corpus of Historical American English</i> (COHA)	1810–2009	400,000,000
<i>Time Magazine Corpus</i> (TIME)	1920s–2000s	100,000,000
British and American		
<i>The Corpus of English Novels</i> (CEN)	1881–1922	26,227,428
British		
<i>Early English Books Online</i> (EEBO)	1470s–1660s	755,000,000
<i>The Proceedings of the Old Bailey</i> (OB)	1674–1913	22,400,000
<i>The Corpus of Late Modern English Texts</i> (CLMET3.0)	1710–1920	34,386,225
<i>British Parliament</i> (Hansard)	1803–2005	1.6 billion

Also providing some examples were *Google Books* and the archive of *The Globe & Mail* newspaper (1844–present). As is noted below, a number of other standard historical corpora provided no examples of this construction.

3.1 Dating of the construction

Early English Books (EEBO)¹⁰ provides a limited number of forms ($n = 43$) but no contracted forms (see (8)).¹¹

10. No examples of *that is not to say* were found in the smaller *Corpus of English Dialogues* 1560–1760 (1,182,690 words) or *Lampeter Corpus of Early Modern Tracts* 1640–1740 (1,193,385 words).

11. The distribution of forms was as follows: *this is not to say* (16 examples); *that is not to say* (5 examples); *which is not to say* (7 examples); and *it is not to say* (15 examples).

- (8) a. *that is not to say* y he hath no regard to me, as though he wer carelesse of the: but by thys maner of speaking we must vnderstand that he loket on y cleannesse of the hart, (EEBO: 1563 Brooke, *The Agreemente of Sondry Places of Scripture*)
- b. *it is not to say* that fayth is a vertue whiche maketh vs righteous in our selues before god ... but we vnderstand that we be iustified by fayth, (EEBO: 1565 Bèze, *A Brief and Piththie Summe of the Christian Faith*)
- c. *this is not to say*, that the doctrine we heare, is holie and good: god shalbe much beholding vnto vs, when wee haue pronounced that his word is worthie to be receiued: (EEBO: 1579 Calvin, *Sermons*, trans. Harmar)
- d. *which is not to say*, that there haue not bin sithence any learned, and eloquent men: but that they are much differing from the former in proprietie, puritie, facilitie, and elegancie of speech; (EEBO: 1594 Leroy, *Of the Interchangeable Course, or Variety of Things*)

Evidence from corpora other than EEBO would suggest rather later origins, with the earliest examples located being *that is not to say* (1825), *this is not to say* (1862), *which is not to say* (1901), and *it is not to say* (1895), as shown below:¹²

- (9) a. *that is not to say*, *that* all ten of the vessels must be finished this year. (1825 Debates in Congress; https://books.google.ca/books?id=Ao4_AQAAMAAJ)
- b. But *that's not to say*, stranger, *that* Nawls aint a good judge. (COHA: 1840 Simms, *Border Beagles*)
- c. *That isn't to say that* it should not be preserved; (1937 G&M 30 Mar.)
- (10) a. *This is not to say that* such persons are not often as deficient in moral discernment as men of large intellectual caliber, (COHA: 1862 *New Englander and Yale Review*)
- b. *This isn't to say that* the Balmy Beaches and the Ottawas started with any overwhelming edge. (1941 G&M 7 Oct.)
- (11) a. they are by far the most abundant visitors, *which is not to say that* long-tongued bees and flies never reach the nectar and transfer pollen, (COHA: 1901 Blanchen, *Wild Flowers*)
- b. Bodily contact ... could make of the sport nothing but a brawl. *Which isn't to say that* hockey doesn't suffer from the same cause, too. (1955 G&M 29 Nov.)

12. No examples of *that is not to say* were found in the text collections *Early English Prose Fiction* (16th and 17th century texts) and *Eighteenth Century Fiction* (texts dating from 1700–1780). ECCO-TCP (2,351 texts from *Eighteenth Century Collections Online*) also supplied no examples, nor did Evans-TCP (6,000 books from the *Evans Early American Imprints*). Only one late example was found in ARCHER (1600–1999).

- (12) a. I am a woman. I follow my brother, whatever he decides. *It is not to say* he is the enemy of persons offending him; only that they have put the division.
(CLMET3.0: 1895 Meredith, *The Amazing Marriage*)
- b. *It's not to say* they don't do something worthwhile besides.
(COHA: 1963 Stohr-Danziger, *Behind the Yellow Door* [play])
- c. *It isn't to say that* his thing comes with a price, because I don't want that to be misunderstood.
(1992 G&M 2 Jun.)

Uncontracted forms predate the contracted forms, but there is not a clear progression from *that* > *this* > *it* > *which* (or any permutation of this order). Excluding the EEBO data, the earliest appearances of the variant *that is not to say* forms fall into three periods (see Table 4).

Table 4. Earliest appearances of the *that is not to say* construction

Period	Instances
early to mid-19th century	<i>that is not to say that</i> (9a), <i>that's not to say that</i> (9b), <i>this is not to say that</i> (10a)
late 19th/early 20th century	<i>that isn't to say that</i> (9c), <i>which is not to say that</i> (11a), <i>it is not to say that</i> (12a)
mid to late 20th century	<i>this isn't to say that</i> (10b), <i>which isn't to say that</i> (11b), <i>it's not to say that</i> (12b), <i>it isn't to say that</i> (12c)

Evidence from COHA shows a steady increase in the frequency of the *that is not to say* construction beginning in the 20th century (see Figure 3).

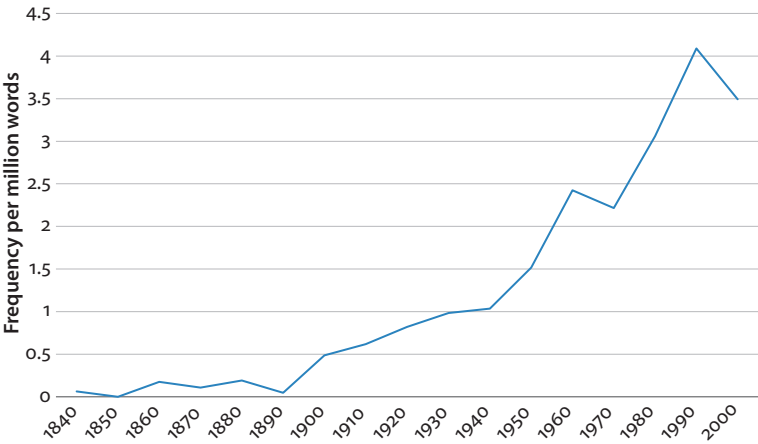


Figure 3. Frequency of the *that is not to say* construction in COHA over time

Because of the indeterminate meaning of the reduced variant *it is not that* in many contexts, where it may either denote non-inference or mean ‘it is not the case that’ (as discussed above in Section 2), it is not possible to trace the development of this form in a precise way. Yet it seems to have a more steady history than *that is not to say*, as seen in the chronological continuity of examples given in (13), all of which are, in my view, interpretable as non-inference markers.

- (13) a. *it is not that* they shoulde bee vnmooueable and vnprofitable when men trouble them: (EEBO: 1574 Calvin, *Sermons* [trans. Golding])
 b. *it is not that* he could not provide for the poore without us, (EEBO: 1633 Balcanquhall, *The Honour of Christian Churches*)
 c. *it is not that* there is any real happiness in what may be acquir’d by means of these divertisements; (EEBO: 1688 *Monsieur Pascall’s thoughts*)
 d. *it is not that* I am so turally prone to it as they, (EEBO: 1694 N.H., *The Ladies Dictionary*)
 e. I would willingly resign my life, *it is not that* I want to save it, (to send me hence is to free me from misery) I only beg your lordship to give me but a lawful chance, (OB: 1756 Trial of John Edgerton Saint, t17560714–22)
 f. *it is not that* I have feared death, no, I have long wished it, (CLMET3.0: 1782 Burney, *Cecilia*)
 g. I fear, every fair, unprejudiced inquirer must in some degree be a sceptic. *It is not that* there are any very staggering arguments against the immortality of man; (CLMET3.0: 1780–96 Burns, *Letters*)

3.2 Formal features of the historical construction

If we compare the form of the *that is not to say construction* in COHA with its form in COCA (see Section 2), we see little change. The presence or absence of complementizer *that* (cf. Tables 1 and 5) changes only slightly, with a decrease in *that* complementation from 69% to 61% and concomitant rise in zero complementation from 31% to 39% over time.¹³

13. In Section 3.3, I present a finer-grained look at the changes in complementizer *that* in COHA.

Table 5. Complementizer *that* in the *that is not to say* construction in COHA

	Ø S		<i>that</i> S		Total
<i>that is not to say</i>	23	(31%)	51	(69%)	74
<i>that isn't to say</i>	2	(50%)	2	(50%)	4
<i>that's not to say</i>	59	(76%)	19	(24%)	78
<i>that ain't to say</i>	1	(100%)	–		1
<i>this is not to say</i>	52	(18%)	243	(82%)	295
<i>this isn't to say</i>	2	(18%)	9	(82%)	11
<i>which is not to say</i>	19	(31%)	43	(69%)	62
<i>which isn't to say</i>	10	(67%)	5	(33%)	15
<i>which ain't to say</i>	1	(100%)	–		1
<i>it is not to say</i>	–		3	(100%)	3
<i>it isn't to say</i>	–		–		–
<i>it's not to say</i>	4	(67%)	2	(33%)	6
Total	173	(31%)	377	(69%)	550 (100%)

One obvious difference is the synchronic preference for *that* over *this* (44% to 38%) in contrast to the diachronic preference for *this* over *that* (56% to 29%). Here, American English seems to be in the lead in replacing *this* by *that* (see Section 2.1 and footnote 2).¹⁴ The relative frequency of *which* and *it* remains stable (see Table 6).

Table 6. Relative frequency of *that*, *this*, *which*, and *it* in the *that is not to say* construction in COHA and COCA

	COHA		COCA	
<i>that</i>	157	(29%)	1,040	(44%)
<i>this</i>	306	(56%)	889	(38%)
<i>which</i>	78	(14%)	338	(14%)
<i>it</i>	9	(2%)	80	(3%)

3.3 Explaining the gaps in the data

The corpus evidence presented above (Section 3.1) points to a significant gap in the data in the 18th century – from early examples in EEBO dating from the 16th and 17th centuries (see Example (8)) to the earliest examples found in other historical corpora dating from the early 19th century (see Examples (9)–(12)). How can this gap be explained? I suggest that the gap in 18th century data for *that is not to say* can be explained by the nature and size of the available corpora, particularly by the preponderance of fiction in many of these corpora.

14. In Brinton (2017: 204) I noticed a similar trend from distal *this* to proximal *that* in *that/this* (*having been*) *said*, perhaps indicating the increasing scope of the construction.

The rarity of the construction is the first issue. Even in contemporary English, we find a relative frequency of only slightly more than four *that is not to say* constructions per million. The large size of EEBO (755 million words) increases the odds of finding such a rare construction compared to the modest size of CLMET3.0 (34 million words) (see Table 3). Moreover, as shown by COHA (see Figure 3), and confirmed by the evidence of Google Books Ngram (see Figure 4) the frequency of *that is not to say* is quite low until the beginning of the 20th century. Thus, the likelihood of encountering this construction in the 18th century remains small.

The second issue is the genre distribution of *that is not to say*. As pointed out in Section 2.1, in Present-day English, the construction has the lowest frequency in Fiction and second lowest in Spoken and is clearly associated with writing-based genres, Newspaper, Magazine, and Academic. While the genre makeup of EEBO is difficult to specify (as it contains all works printed during the period), most of the texts could best be described as non-fiction prose, with some narrative, but not what we would consider fiction or the novel. It is well accepted that the novel, as we know it, did not arise in English until the 18th century or perhaps late 17th century (e.g. Watt 2001/1957). The non-fiction prose of EEBO would thus be a rich source for *that is not to say*, based on its genre distribution in Present-day English. However, the two Chadwyck-Healey fiction collections (see footnote 12) and CEN, all of which consist entirely of fiction, would not be expected to yield examples of this construction.¹⁵ OB, which in many ways approximates spoken English, also unsurprisingly contains only three examples, all from 1900.¹⁶ In CLMET3.0, narrative fiction and drama constitute nearly one half of the corpus (49.6%).¹⁷ Drama is included in the calculation as it has been found to be closest to spoken in historical texts dating from 1600–1700 (Culpeper & Kytö 2000). If narrative non-fiction is also included, the proportion of the corpus in which we would not expect to find *that is not to say* rises to 63.8%. COHA has no newspaper data until the 1860s, and even when it begins to appear, it supplies less than 1.5 million words/decade until the 1920s (constituting only 6–7% of the corpus in the 19th century). More importantly, fiction predominates at all periods in COHA, ranging from 48–55% of the corpus (averaging 51%).¹⁸

15. I cannot know exactly what texts have been included in ECCO-TCP and Evans-TCP (see footnote 12), though presumably there is a high proportion of fiction.

16. The lack of 19th century examples in Hansard (of the 4,673 examples, all but one are post 1900) remains a puzzle. It may be, as one reviewer suggests, due to the strong influence of several scribes or to the fact that before 1909, Hansard did not represent exact transcriptions of the parliamentary proceedings.

17. See <https://perswww.kuleuven.be/~u0044428/>.

18. See <<https://corpus.byu.edu/coha.asp>>.

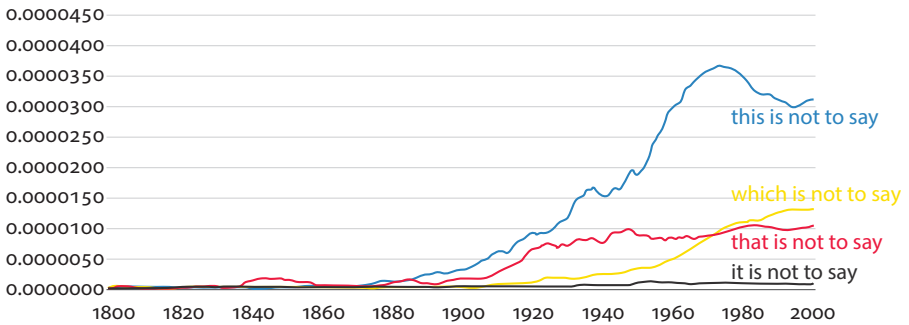


Figure 4. Frequency of all forms of *that is not to say* in Google Books over time (Google Ngram)

3.4 Grammaticalization of *that is not to say*

As argued above, *that is not to say* can be understood as a discursively secondary and grammatical form, which Boye and Harder (2007, 2012) argue arises through grammaticalization. We see in *that is not to say* many of the characteristics typically associated with grammaticalization, especially the grammaticalization of comment clauses (see, e.g., Brinton 2008: 58–61, 241–245). The construction has undergone decategorialization (Hopper 1991) from a matrix clause, or full complement-taking predicate, to a comment clause; this shift is characterized by Boye and Harder as the change of a lexical and primary CTP to a grammatical and secondary CTP involving reanalysis on both the usage and structural level (2007: 590). With the increase of contracted forms, we see coalescence, or the reduction of phonological segments, typical of grammaticalization. Formal freezing also occurs: adverbial modification is ruled out and there is limited variation in the subject form, especially in American English, which appears to be moving toward the fixing of *that* as the preferred subject pronoun. Desemanticization of *say* from the meaning ‘utter in words’ to the bleached meaning ‘indicate’ occurs, though traces of the original meaning persist (Hopper’s principle of “persistence” (1991: 2)): in the broadest sense *that is not to say* refers to something that has been “said” in the previous discourse. More importantly, *that is not to say* follows the correlated semantic pathways identified by Traugott and Dasher (2002: 40, 281) involving a shift from content to procedural meaning. The acquisition of a meaning of non-inference represents pragmatic expansion or strengthening. It also follows the pathway of (inter)subjectification; that is, from the content meaning ‘that should not be uttered’, it acquires the subjective meaning ‘I am not implying X’ and then the intersubjective meaning ‘you should not infer X’. Finally, the rise of newer non-inference forms such as *I’m just saying* or *all I’m saying* (see Brinton 2017: 209–210) points to the principle of layering in grammaticalization identified by Hopper (1991).

Two factors related to grammaticalization warrant further examination: contraction and *that*-deletion. The first bears on the issue of coalescence, while the second bears on the grammatical and secondary status of *that is not to say*.

The use of contracted forms in the *that is not to say* construction rises from under 10% to over 50% in COHA over time (see Figure 5); this increase could thus be seen as strong evidence of coalescence. However, the increasing number of contracted forms may be the result of *colloquialization*, changes in the norms of written English to more resemble those of oral speech. Leech et al. (2009: 128–129, 240–241) consider contraction as a “paradigm case of colloquialization” and see contractions as having “increased enormously” over sixty years; only the academic genre seems resistant to this increase. López-Couso (2007) in a more nuanced study of contractions in Late Modern English finds a significant increase in contracted *have* and *be* in 18th and 19th century British English, with a preference for auxiliary enclisis over negative enclisis. This is especially true with present tense forms of *be* and pronoun subjects (López-Couso 2007: 311, 316–319; see also Biber et al. 1999: 1128–1132). Thus, the fact that *that’s not to say* and *it’s not to say* are strongly favored over *that isn’t to say* and *it isn’t to say* in the corpora examined, while *this isn’t to say* and *which isn’t to say* are very low frequency (see Tables 1, 2, and 5) is entirely consistent with general patterns of contraction.

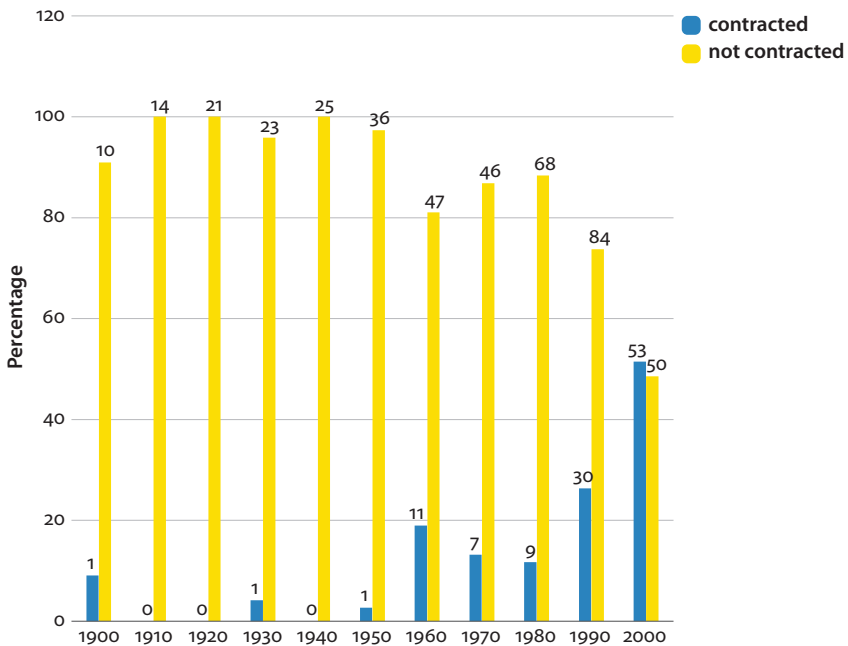


Figure 5. Contracted and uncontracted forms of the *that is not to say* construction over time in COHA

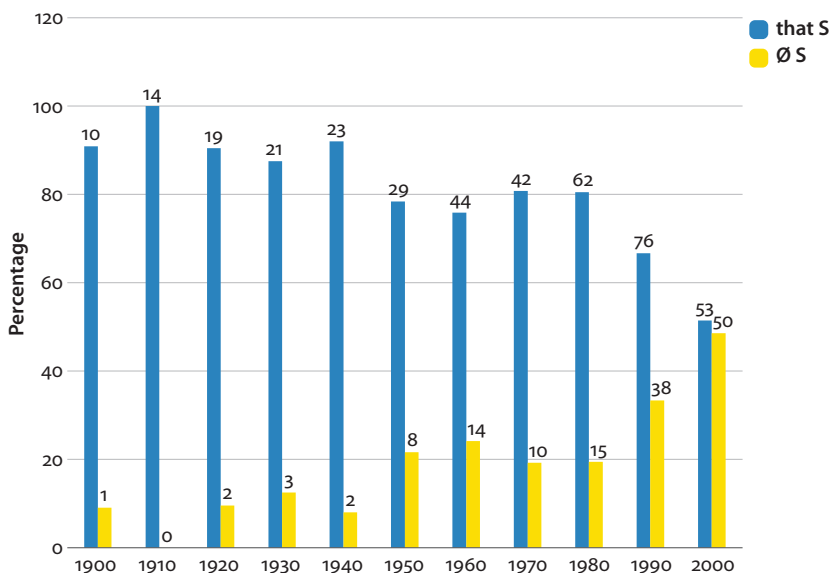


Figure 6. *That/Ø* complements with the *that is not to say* construction over time in COHA

The deletion of *that* in the *that is not to say* construction also rises in COHA from 9% to 49% (see Figure 6), thus suggesting the evolution of the construction from a matrix clause to a discourse secondary, grammaticalized form. However, caution must be exercised for three reasons. First of all, there is the low number of zero-complements found before the mid-20th century in COHA. Secondly, we should take into account the synchronic debate, as discussed in Section 2.2.2 earlier, concerning the significance of *that*-deletion or retention for determining the syntactic and discourse status of the CTP. Third is the common understanding that *that* deletion is a consequence of colloquialization and has increased over time as written English has become more speech-like. While this is in part true, especially as we see the correlation of higher rates of *that*-deletion with more colloquial contracted forms (Tables 1, 2, and 5), we must recognize that *that*-deletion has taken a remarkably uneven course over time, disrupted perhaps by proscriptions concerning the deletion of *that*. Compiling a meta-analysis of four studies of *that*-deletion based on three historical corpora, Torres Cacoullós and Walker (2009: 5) show that zero-complements in fact reached a peak in the mid-17th century, with a steady decline thereafter. Rissanen (1991) too finds that the omission of *that* was most frequent in the second half of the 16th and part of early 17th century, with *that* complementizers making a come-back in the ‘norm-loving’ 18th century (see also Finegan & Biber 1995).

A final question here is the apparent lack of the purely lexical and primary CTP stage (Boye & Harder 2007) in the historical data examined, and the absence of “divergence” (Hopper 1991: 22), or the continued existence of the original lexical stage. *That is not to say (that)* seems to be already fully grammaticalized in the earliest 16th century examples in EEBO. A fuller accounting of the history of this form would thus require an examination of its Middle English antecedents, if any. While the *Middle English Dictionary* (MED, s.v. *that* (pron.), def. (1a)(d)) lists the existence of the positive form in a number of variants (*that is to seien (knouen, menen, witen)*), the negative form is unrecorded.

4. Conclusions

Synchronically, *that is not to say* is different in form and function as well as history from its positive counterpart, the appositive marker *that is to say*. It introduces an inference which should not be drawn from the previous discourse and is often followed by a negative that explicitly denies this inference. Variants are formed with the distal and proximal demonstratives, and with *which* and *it*, although *which* and *it* exist only as minor variants; American English would appear to be leading the way in preferring *that* to *this*. The construction typically includes the complementizer *that*, although it may be deleted, especially in variants containing contracted forms. *That is not to say* meets many of the criteria of a grammatical, discursively secondary CTP (complement-taking predicate) in Boye and Harder’s (2007) sense. But as with other such constructions (e.g. *I think, it seems*), the presence or absence of *that* is not determinative of its lexical or grammatical status. In present-day corpora, the construction is rare in fiction and its frequency is also low in spoken language; it is thus typical of written genres. The existence of triple negatives (with both positive and negative interpretations) in web data possibly represents an expansion of the construction, but could also be construed as a misnegation (Lieberman 2012).

Diachronically, the construction appears as a non-inference introductory formula already in the 16th and 17th centuries in *Early English Books Online*. However, examples cannot be found in the 18th century, becoming more common only in the mid-19th century, with contracted forms appearing later. The gap in the data most likely results from the predominantly fiction/drama composition of *The Corpus of Late Modern English Texts 3.0* and *The Corpus of Historical American English*; in fact, newspaper collections from the period provide some of the earliest examples of *that is not to say* constructions. In many respects, the development of *that is not to say* can be understood as a case of grammaticalization, showing decategorialization, coalescence and freezing, desemanticization, (inter)

subjectification, pragmatic enrichment, and layering. However, two explicit signs of grammaticalization (loss of complementizer *that* and the rise of contracted forms) are not conclusive, being perhaps attributable to colloquialization, or the trend of written English towards the norms of spoken English. The existence of antecedents of *that is not to say* in Middle English requires further study in order to establish the lexical stage of the construction.

References

Corpora and text collections

- ARCHER = *A Representative Corpus of Historical English Registers*. Version 3.2. 1990–1993/2002/2007/2010/2013/2016. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. <<https://cqpweb.lancs.ac.uk>>
- BYU-BNC = Davies, Mark. 2004–. BYU-BNC (Based on the British National Corpus from Oxford University Press). <<https://www.english-corpora.org/bnc/>>
- A Corpus of English Dialogues 1560–1760*. 2006. Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University). <<http://www.engelska.uu.se/forskning/engelska-spraket/elektroniska-resurser/a-corpus>>
- CEN = *The Corpus of English Novels*. De Smit, Hendrik (comp.). <<https://perswww.kuleuven.be/~u0044428/cen.htm>>
- CLMET3.0 = *The Corpus of Late Modern English Texts, version 3.0*. De Smet, Hendrik, Diller, Hans-Jürgen & Tyrkkö, Jukka (comps). <https://perswww.kuleuven.be/~u0044428/cl-met3_0.htm>
- COCA = Davies, Mark. 2008–. *The Corpus of Contemporary American English: 520 million words, 1990–2015*. <<https://www.english-corpora.org/coca/>>
- COHA = Davies, Mark. 2010–. *The Corpus of Historical American English: 400 million words, 1810–2009*. <<https://www.english-corpora.org/coha/>>
- Early English Prose Fiction*. 1997–2015. Ed. by Holger Klein et al. Chadwyck-Healey. ProQuest LLC. <http://collections.chadwyck.com/home/home_eepf.jsp>
- ECCO-TCP = *Eighteenth Century Collections Online* <<https://www.textcreationpartnership.org/tcp-ecco/>>
- EEBO = Davies, Mark. 2017. *Early English Books Online*. Part of the SAMUELS project. <<https://www.english-corpora.org/eebo/>>
- Eighteenth-century Fiction (1700–1780)*. 1996–2015. Chadwyck-Healey (ProQuest Company). <http://collections.chadwyck.com/marketing/home_c18f.jsp>
- Evans-TCP = *Evans Early American Imprints*. <<https://www.textcreationpartnership.org/tcp-evans/>>
- G&M = *Globe and Mail* 1844–present. ProQuest Historical Newspapers. <<https://www.proquest.com/about/news/2012/The-Globe-and-Mail-Joins-ProQuest-Historical-Newspapers.html>>
- Google Books. <<https://books.google.com>>

- Hansard = Alexander, Marc & Davies, Mark. 2015–. *British Parliament – Hansard 1803–2005*. Part of the SAMUELS project. <<https://www.english-corpora.org/hansard/>>
- The Lampeter Corpus of Early Modern English Tracts*. 1998. Compiled by Josef Schmied, Claudia Claridge & Rainer Siemund (University of Chemnitz). <<https://www.tu-chemnitz.de/phil/english/sections/linguist/real/independent/lampeter/lamphome.htm>>
- OB = *The Old Bailey Proceedings Online, 1674–1913*. Hitchcock, Tim, Shoemaker, Robert, Emsley, Clive, Howard, Sharon & McLaughlin, Jamie et al. (eds). <www.oldbaileyonline.org> version 7.0, March 2012.
- TIME = Davies, Mark. 2007–. *TIME Magazine Corpus: 100 million words, 1920s–2000s*. <www.english-corpora.org/time/>

Other references

- Aijmer, Karin. 1997. *I think* – an English modal particle. In *Modality in Germanic Languages: Historical and Comparative Perspectives*, Toril Swan & Olaf Jansen Westvik (eds), 1–47. Berlin: Mouton de Gruyter.
- Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education. <https://doi.org/10.1177/007542420225290>
- Boye, Kasper & Harder, Peter. 2007. Complement-taking predicates: Usage and linguistic structure. *Studies in Language* 31(3): 569–606. <https://doi.org/10.1075/sl.31.3.03boy>
- Boye, Kasper & Harder, Peter. 2012. A usage-based theory of grammatical status and grammaticalization. *Language* 88(1): 1–44. <https://doi.org/10.1353/lan.2012.0020>
- Brinton, Laurel J. 2008. *The Comment Clause in English: Syntactic Origins and Pragmatic Development*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511551789>
- Brinton, Laurel J. 2017. *The Evolution of Pragmatic Markers in English: Pathways of Change*. Cambridge: CUP. <https://doi.org/10.1017/9781316416013.002>
- Briscoe, Ted. 2000. Interview of Gerald Gazdar. On Generalized Phrase Structure Grammar. 3 November 2000, Cambridge, UK <<https://nlp.fi.muni.cz/~xjakub/briscoe-gazdar/gpsg.html#SECTION00040000000000000000>> (1 May 2018).
- Culpeper, Jonathan & Kytö, Merja. 2000. Data in pragmatics: Spoken interaction (re)cast as writing. *Journal of Historical Pragmatics* 1(2): 175–199. <https://doi.org/10.1075/jhp.1.2.03cul>
- Dehé, Nicole & Wichmann, Anne. 2010. Sentence-initial *I think (that)* and *I believe (that)*: Prosodic evidence for the use of main clause, comment clause and discourse marker. *Studies in Language* 34(1): 36–74. <https://doi.org/10.1075/sl.34.1.02deh>
- Finegan, Edward & Biber, Douglas. 1995. *That* and zero complementisers in Late Modern English: Exploring ARCHER from 1650–1990. *The Verb in Contemporary English*, Bas Aarts & Charles F. Meyer (eds), 241–257. Cambridge: CUP.
- Google Books Ngram Viewer. <<http://books.google.com/ngrams>>. See Michel et al. 2011.
- Goldberg, Adele E. & van der Auwera, Johan. 2012. This is to count as a construction. *Folia Linguistica Historica* 46(1): 109–132. <https://doi.org/10.1515/flin.2012.4>
- Hopper, Paul J. 1991. On some principles of grammaticization. In *Approaches to Grammaticalization, Vol. 1: Theoretical and Methodological Issues* [Typological Studies in Language 19], Elizabeth Closs Traugott & Bernd Heine (eds), 17–35. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.19.1.04hop>
- Hopper, Paul J. & Thompson, Sandra A. 2008. Projectability and clause combining in interaction. In *Crosslinguistic Studies of Clause Combining: The Multifunctionality of Conjunctions* [Typological Studies in Language 80], Ritva Laury (ed.), 99–123. Amsterdam: John Benjamins.

- <https://doi.org/10.1075/tsl.80.06hop>
- Huddleston, Rodney & Pullum, Geoffrey K. et al. 2002. *The Cambridge Grammar of the English Language*. Cambridge: CUP. <https://doi.org/10.1017/9781316423530>
- Kaltenböck, Gunther. 2011. Explaining diverging evidence: The case of clause-initial *I think*. In *Converging Evidence: Methodological and Theoretical Issues for Linguistic Research* [Human Cognitive Processing 33], Doris Schönefeld (ed.), 81–112. Amsterdam: John Benjamins. <https://doi.org/10.1075/hcp.33.07kal>
- Kearns, Kate. 2007. Epistemic verbs and zero complementizer. *English Language and Linguistics* 11(3): 475–505. <https://doi.org/10.1017/S1360674307002353>
- Koivisto-Alanko, Päivi & Rissanen, Matti. 2002. *We give you to wit*: Semantics and grammaticalisation of the verb *wit* in the history of English. In *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen*, Helena Raumolin-Brunberg, Minna Nevala, Arja Nurmi & Matti Rissanen (eds), 13–32. Helsinki: Société Néophilologique.
- Leech, Geoffrey, Hundt, Marianne, Mair, Christian & Smith, Nicholas. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511642210>
- Liberman, Mark. 2012. This is not to say that I don't think that it isn't illogical. *Language Log*, 3 November 2012, <languagelog.ldc.upenn.edu/nll/?p=4292> (10 September 2018).
- López-Couso, María José. 2007. Auxiliary and negative cliticisation in Late Modern English. In *'Of varying language and opposing creeds': New Insights into Late Modern English*, Javier Pérez-Guerra, Dolores González-Álvarez & Jorge L. Bueno Alonso (eds), 301–323. Bern: Peter Lang.
- Longman Dictionary of Contemporary English*. <<https://www.ldoceonline.com/dictionary/that-is-not-to-say>> (10 September 2018).
- Macmillan Dictionary*. <<https://www.macmillandictionary.com/dictionary/british/that-s-not-to-say-that>> (10 September 2018).
- MED = *Middle English Dictionary*. 1952–2001. Hans Kurath et al. (eds). Ann Arbor, MI: Michigan University Press.
- Merriam-Webster Dictionary*. <<https://www.merriam-webster.com/dictionary/that%27s%20not%20to%20say>> (10 September 2018).
- Michel, Jean-Baptiste, Shen, Yuan Kui, Presser Aiden, Aviva, Veres, Adrian, Gray, Matthew K., The Google Books Team, Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A., & Lieberman Aiden, Erez. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014), 176–182. <https://doi.org/10.1126/science.1199644>
- Moore, Colette. 2006. The use of *videlicet* in Early Modern English slander depositions. *Journal of Historical Pragmatics* 7(2): 245–263. <https://doi.org/10.1075/jhp.7.2.05moo>
- OED = *Oxford English Dictionary Online*. 2000–. Michael Proffitt (ed.). 3rd edn. Oxford: OUP. <<http://www.oed.com/>> (30 May 2019).
- Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Rissanen, Matti. 1991. On the history of *that*/zero of object clause links in English. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Karin Aijmer & Bengt Altenberg (eds), 272–289. London: Longman.
- Schilperood, Joost & Verhagen, Arie. 1998. Conceptual dependency and the clausal structure of discourse. In *Discourse and Cognition: Bridging the Gap*, Jean-Pierre König (ed.), 141–163. Stanford, CA: CSLI.

- Thompson, Sandra A. 2002. "Object complements" and conversation: Towards a realistic account. *Studies in Language* 26(1): 125–164. <https://doi.org/10.1075/sl.26.1.05tho>
- Thompson, Sandra A. & Mulac, Anthony. 1991. A quantitative perspective on the grammaticalization of epistemic parentheticals in English. In *Approaches to Grammaticalization, Vol. 2: Types of Grammatical Markers* [Typological Studies in Language 19], Elizabeth Closs Traugott & Bernd Heine (eds), 313–329. Amsterdam: John Benjamins. <https://doi.org/10.1075/tsl.19.2.16tho>
- Torres Cacoullous, Rena & Walker, James A. 2009. On the persistence of grammar in discourse formulas: A variationist study of *that*. *Linguistics* 47(1): 1–43. <https://doi.org/10.1515/LING.2009.001>
- Traugott, Elizabeth Closs & Dasher, Richard B. 2002. *Regularity in Semantic Change*. Cambridge: CUP. <https://doi.org/10.1017/CBO9780511486500>
- Verhagen, Arie. 2001. Subordination and discourse segmentation revisited, or: Why matrix clauses may be more dependent than complements. In *Text Representation: Linguistic and Psychological Aspects* [Human Cognitive Processing 8], Ted J. M. Sanders, Joost Schilperood & Wilbert Spooren (eds), 337–357. Amsterdam: John Benjamins. <https://doi.org/10.1075/hcp.8.18ver>
- Watt, Ian. 2001/1957. *The Rise of the Novel: Studies in Defoe, Richardson and Fielding*. Berkeley, CA: University of California Press.

Changes in transitivity and reflexive uses of *sit* (*me/myself down*) in Early and Late Modern English

Turo Vartiainen and Mikko Höglund
University of Helsinki / Stockholm University

This chapter seeks to establish if the Transitivity Hypothesis (Hopper & Thompson 1980) can explain the variation in the use of two reflexive strategies with the verb *sit* in Early Modern English (e.g. *I sat me down/I sat myself down*) and the verb's subsequent transitivity (e.g. *he sat me down*). By studying data from large historical corpora, we will re-evaluate the results of earlier research and establish why *sit* continued to be used with the simple reflexive strategy (i.e. with object pronouns) until the Late Modern period. In our analysis of the transitivity of *sit* (*down*), we focus on both micro-level semantic and syntactic factors and more general developments that have supported the transitivity of verbs in Late Modern English.

1. Introduction

The verbal system of English has undergone many changes in the course of history from the perspective of transitivity. For example, it is well-documented that many verbs that are typically intransitive in Present-day English (henceforth PDE) could also be expressed in reflexive constructions in earlier periods (e.g. Mitchell 1985: 113; Mustanoja 1960: 100; van Gelderen 2000). Even today, some reflexive patterns persist in dialectal usage and fossilized expressions (Peitsara 1997: 323, 349), and their archaic nature is occasionally exploited by fiction writers, who use them to index the speech of historical characters. For example, although the reflexive pattern with *fear* (e.g. *I fear me that I cannot come*) had largely been replaced by an intransitive pattern in Early Modern English (Peitsara 1997: 325–328), we can still find it used in 20th-century corpus data, particularly in historical novels (as in (1)). Likewise, the reflexive use of *lie* (which is often used interchangeably with *lay*, as in (2)) still enjoys marginal currency in PDE, probably in part because of

a popular children's bedtime prayer which dates back to the 18th century (*When I lay me down to sleep...*).

- (1) *I fear me* both are false. (COHA:¹ Fiction, 1955)
 (2) I would love to see a lot of things happen between now and the time *I lay me down to rest*. (COHA: Non-fiction, 2009)

The semi-fossilized nature of these patterns becomes evident when we consider their limited potential for variation. For example, the patterns exemplified in (1) and (2) are largely restricted to being used in the first person singular. So, while *I lay me down* is well-attested in COHA, there are no instances of *he lay him down*, and the most recent token of *she lay her down* dates back to 1834 (see also Keenan 2002: 348). For most verbs, however, the earlier reflexive patterns are no longer available. For instance, motion verbs like *come*, *go*, and *ride*, which could be used in reflexive constructions in Old and Middle English (e.g. Huber 2017; Visser 1963: 321), are invariably intransitive in PDE, and verbs of posture like *rest* and *stand* have likewise shed the earlier reflexive pattern in favour of the intransitive pattern (as in (3) and (4)).

- (3) Þe king *him* *rod* an huntunge
 the king he-REFL ride-3SG-PRET to hunting
 'The king rode to the hunt.' (Helsinki Corpus: *King Horn*, 1250–1350)
 (4) Þa *stod* *him* *up* Walwain
 then stand-3SG-PRET he-REFL up Walwain
 'Then Walwain stood up.' (Helsinki Corpus: *Layamon*, 1150–1250)

In addition to this macro-trend which has favoured intransitive uses over reflexive ones, the pronouns that have been used to express reflexivity have changed in the course of history. The earlier simple reflexives, exemplified in (1)–(4), gradually gave way to the SELF-paradigm, which is also the paradigm still in use in PDE. This process started in Middle English and continued well into the Early Modern period. Changes in reflexive strategies have been documented in detail in, for instance, Peitsara (1997) and van Gelderen (2000), but as both of these studies were done at a time when linguistic corpora and databases were much smaller than today, their analyses are often based on very low token frequencies and therefore cannot be considered conclusive in every respect.² Indeed, with new historical databases at our disposal, we are now in a position to complement the results obtained in earlier

1. COHA = *The Corpus of Historical American English* (Davies 2010).

2. For example, Peitsara (1997), using the Helsinki Corpus, was only able to find 20 tokens of verbs of posture and change of posture (*lie*, *rest*, *sit*, *stand*, *lay*, *lift*, *raise*, *rouse*) for the entire

research with much larger datasets and shed more light on the variation between different reflexive strategies in earlier forms of English.

Our study focuses on a single verb that was one of the last verbs to hold on to the simple reflexive pattern: *sit*. We will study two research questions related to the use of *sit* in the Early and Late Modern periods. First, we will examine the linguistic contexts that affected the variation between the use of simple reflexives (e.g. *I sit me down*) and SELF-reflexives (e.g. *I sit myself down*) in the Early Modern period. Our results show that in order to understand the variation between the two reflexive strategies we must pay close attention to what kind of event the clause describes aspectually. Our second question focuses on the development of a transitive pattern that has recently become increasingly popular, at least in American English (e.g. *he sat me down*). The history of this usage dates back to the 18th century, and our analysis reveals that the innovation was supported by an overlap between the uses of *sit* and *set*, on the one hand, and by a semantic preference according to which *set* was not well suited for descriptions of events where the potency of the agent was realized only indirectly. We will explain both the variation in the simple and SELF-reflexives and the transitivization of the *sit N down* pattern in terms of the Transitivity Hypothesis, which was first formulated in Hopper and Thompson (1980).

This chapter is organized as follows. Section 2 provides some necessary background to the case studies discussed in this chapter. We will first summarize some of the relevant literature on reflexivity and the history of reflexive verbs in English. This discussion is followed by a description of the Transitivity Hypothesis (Hopper & Thompson 1980), which will be the framework within which we interpret and analyse our data, as well as definitions of the aspectual categories that will be central to our analysis. Section 3 provides a description of the corpora and databases used in this study as well as a discussion on data collection and categorization. In Section 4, we present the results of our case studies, and Section 5 concludes the chapter with a discussion of the main findings and some suggestions for future research.

2. Theoretical background and reflexivity in the history of English

2.1 Defining reflexivity

Reflexivity can be defined as a coreferential relation between the two arguments of a verb (see e.g. Crystal 2008: 408; Lyons 1968: 361). This relationship can be conventionally expressed by coindexing the arguments, as in (5a), while the differing

Early Modern period (1500–1710). To compare, our data for *sit* alone includes 1,550 tokens (1500–1700).

indices in (5b) indicate that the referents of the pronouns should be understood as different entities.

- (5) a. He_i hit himself_i.
 b. He_i hit him_j.

Reflexivity can be encoded in language in different ways. Reuland (2000: 14) makes a distinction between pronominal marking (e.g. Dutch *hem*, Icelandic *hann*), simplex anaphors (e.g. Dutch *zich*, Icelandic *sig*) and complex anaphors (e.g. Dutch *zichzelf*, English *himself*). In the first group, reflexivity is expressed with pronouns that have other (primary) functions in the language, for instance to express object case. In the simplex group, the pronoun is only used to express reflexivity, while in the complex group the reflexive form consists of a pronoun followed by another element (e.g. *himself*). The difference between pronominal marking and the other two strategies can also be discussed in terms of a difference between pragmatic marking (pronominal strategy) and grammatical marking (simplex and complex strategies) (Ariel 2012: 41). These strategies are not mutually exclusive; rather, a language can make use of all these strategies to varying degrees. In PDE, the SELF-strategy is the dominant strategy, but the pronominal strategy can be used, for example, with some oblique constructions, as in (6) (from Ariel 2012: 41; see also van Gelderen 2000: 111).

- (6) He_i had no spots on him_i.

Furthermore, there are verbs whose meaning includes a clear element of self-directedness, such as *wash* and *shave* (cf. Ito 1978), which are sometimes called “semi-reflexives” in the literature (e.g. Quirk et al. 1985: 358). With such verbs, the reflexive pronoun is optional, and both *he shaved himself* and *he shaved* are perfectly grammatical. In other words, these constructions allow both implicit (or covert) and explicit (or overt) expressions of reflexivity (Lyons 1968: 361–362).

2.2 Reflexive strategies in the history of English

In English, the reflexive paradigm has undergone a change from marking reflexivity with simple accusative and dative pronouns to having *self*-marked reflexives (i.e. from the simple strategy to the SELF-strategy). This development was gradual, and it involved several steps that took place at different times for different verb types. The general line of development is relatively clear, however. First, in Old English only the simple strategy was used, and *self* was an adjectival word that could follow the pronoun to indicate emphasis (as in (7)).

- (7) Judas *hine sylfne aheng*.³
 Judas 3SG.ACC self-MASC.SG.ACC ahang-RET-INDIC.3SG
 ‘Judas hanged himself.’³ (Ælfric’s Homilies II, 250, 15)

Gradually, during the Middle English period, sound changes led to the reinterpretation of pronoun+*self* combinations, so that the dative pronouns in structures like *me self* were reanalysed as genitives (Mustanoja 1960: 146). The adjectival *self*, on the other hand, was reanalysed as a noun, which in turn paved the way for the further grammaticalization of the pronoun+*self* forms into reflexive pronouns (see e.g. Keenan (2002) for a more detailed account of this process).

In PDE, the *self*-forms are locally bound whereas the simple pronouns cannot be locally bound, that is, they are locally free (Keenan 2002: 329). This means that in (8a) *himself* is coreferential with *the king*, and in (8b) *him* is coreferential with *John*.

- (8) a. John denied that the king_i sat himself_i down.
 b. John_i denied that the king sat him_i down.

However, in Old English the simple strategy was the only way to mark reflexivity, and so both local and non-local binding relations were expressed with simple pronouns. Consequently, a sentence like (8b) would have been ambiguous regarding coreferentiality. The creation of the reflexive *self*-pronouns, starting with the use of *self* as a contrastive device in Old English and its merging with the dative and accusative pronouns in c. 13th century (Keenan 2002: 333–337), provided another strategy to express coreferential relations, and by the end of the 16th century, the *self*-forms had lost their obligatory contrastive function and took over local binding from the simple pronouns. According to Keenan (2002: 346–347), the shift in binding in the 1500s was largely due to the functional similarity of the *self*-pronouns and simple pronouns in reflexive expressions: a locally bound *self*-form was synonymous with a locally bound simple pronoun. The solution that presented itself was that the *self*-forms in object position came to require local antecedents, and the simple pronouns started to reject these – which is also the situation in PDE.

As shown in Peitsara (1997), the older simple strategy remained more common in English until the end of the Middle English period (1420–1500). However, the situation changed rapidly in the 16th century: while in 1420–1500, 72% of all reflexives in Peitsara’s data were expressed with the simple strategy, in 1500–1570, the corresponding proportion is as low as 33%. In 1640–1710, the simple strategy

3. See the OED for this obsolete verb meaning ‘hang’.

only accounts for 2% of all reflexive uses (Peitsara 1997: 288), showing that the older strategy practically fell out of use within a period of approximately 200 years. However, it should be noted that this overall development conceals some interesting variation. For instance, while many verbs of psychological events, such as *hate*, *love*, and *pity*, are not attested with the simple strategy after 1500 in the Helsinki Corpus, verbs of posture, such as *sit* and *lie*, could still be used with simple pronouns in the late 17th century (Peitsara 1997: 322, 326). Furthermore, psych-verbs that express more involuntary actions, such as *fear*, *grieve* and *repent*, are attested with the older simple strategy as late as 1570–1640, confirming that the survival of the older pattern was in part dependent on verb type (see also Ito 1978 for a discussion of the syntactic versus semantic classification of reflexives).

There is also evidence that the diffusion of the SELF-strategy did not proceed at a steady rate in all genres and contexts of use. First, Peitsara (1997: 291–294) shows that the simple strategy persisted longer in oral genres in Early Modern English (see also Spies 1897: 157). Second, van Gelderen (2000: 79) demonstrates that in Early Middle English (e.g. Layamon's *Brut*), reflexive *self*-forms were particularly often used with third person pronouns and with prepositional objects. In Chaucer's texts, on the other hand, first person forms stand out against second and third person forms by continuing to favour the earlier simple strategy. The importance of third person forms in the diffusion of the SELF-strategy can also be seen in the spelling of Early Modern texts. In Shakespeare's First Folio, for example, forms like *himself*, *himsel*, and *him-sel* are used 428 times, but there are no examples of the form written separately as *him self(e)*. By contrast, there are only two forms of *myself* but hundreds of *my self* (van Gelderen 2000: 113). So, from very early on we see a split in the person system that has persisted in some of the archaic forms that were discussed in Section 1 (e.g. *I lay me down to rest*).

2.3 The Transitivity Hypothesis in the context of this study

The Transitivity Hypothesis was first proposed by Paul Hopper and Sandra Thompson in an article published in *Language* (1980). In short, the hypothesis predicts that syntactic transitivity has semantic/pragmatic correlates, so that if a language expresses one meaning relationship with transitive syntax and another one with non-transitive syntax, the split should follow a principled semantic motivation predicted by the Transitivity Hypothesis. In their article, Hopper and Thompson (1980: 252) list ten relevant features that affect the overall transitivity of a clause (Table 1).

What the Transitivity Hypothesis suggests, then, is that there is a cline of transitivity and that languages can organize the description of various kinds of events in different ways insofar as they follow the implications of the transitivity parameters

Table 1. Transitivity parameters according to the Transitivity Hypothesis

		High	Low
A.	Participants	2 or more participants, A and O	1 participant
B.	Kinesis	action	non-action
C.	Aspect	telic	atelic
D.	Punctuality	punctual	non-punctual
E.	Volitionality	volitional	non-volitional
F.	Affirmation	affirmative	negative
G.	Mode	realis	irrealis
H.	Agency	A high in potency	A low in potency
I.	Affectedness of O	O totally affected	O not affected
J.	Individuation of O	O highly individuated	O non-individuated

listed in Table 1. For instance, while in English a clause like *I like beer* (which is relatively low in transitivity in terms of parameters B, D, E, H and I) is expressed with transitive syntax, in Spanish the same meaning is expressed with only one core argument, and the experiencer is encoded in the dative case: *me gusta la cerveza* (Hopper & Thompson 1980: 254). In the history of English, we find a similar impersonal construction (see e.g. Möhlig-Falke 2012; Palander-Collin 1999), showing that a low degree of semantic transitivity used to be connected with non-transitive syntax also in English, but in PDE all such meanings are expressed with syntactically transitive clauses.

The Transitivity Hypothesis was designed to account for the various ways in which languages encode meanings on the clausal level, and it has mainly been used in typological studies (see e.g. Næss 2007; Nordlinger 2011). However, the hypothesis has received relatively little attention as an explanation for diachronic developments. The reason for this is not entirely clear to us; if a hypothesis is able to explain synchronic variation both within individual languages and cross-linguistically, it stands to reason that it could offer new insights into language change as well. When it comes to expressions of reflexivity in English, the change from the simple to the SELF-strategy could be interpreted as being motivated by the low degree of semantic transitivity expressed by reflexive clauses (insofar as the development is perceived as a change from accusative pronouns, which are used to mark core arguments, to *self*-reflexives).⁴ Furthermore, as we have already seen, some verbs actually retained the older simple strategy long after most verbs had adopted the SELF-strategy. Because the distinction between dative and accusative pronouns had

4. Hopper and Thompson (1980: 277–278) point out that in many languages, reflexives have an intermediate status between one-participant and two-participant clauses. See also Mondorf (2016: 83–84), who extends this analysis to the dummy object *it*.

been lost in Early Modern English, all reflexive clauses with simple pronouns can only be analysed as syntactically transitive in EModE.

What we are interested in establishing in this chapter is whether the Transitivity Hypothesis might offer insight into the fact that *sit* was one of the only verbs to retain the simple reflexive strategy for a prolonged period of time. Why would this be? Was *sit* used in contexts that were high in semantic transitivity and which would therefore support transitive syntax? Of course, when studying *sit*, the agent in the described event is typically high in potency and acts volitionally, which means that most of the clauses in our data are high in transitivity according to these two parameters. However, there are other transitivity parameters that may show variation. More specifically, clauses with *sit* may express either action or non-action (e.g. *he sits down* vs. *he sits in council*; Parameter B, Kinesis) and be aspectually telic or atelic (e.g. *he sat down* vs. *he is sitting in the garden*; Parameter C, Aspect). These parameters are in principle independent from each other. For instance, according to our analysis *he sat in the garden for three hours* would be an aspectually telic description of a non-action. In our data, however, the parameters show nearly perfect overlap: there are only seven aspectually telic clauses that express non-action in the data. Furthermore, all descriptions of atelic eventualities represent non-action. Because of this substantial overlap, we will report our results in Section 4.2 only in terms of telicity, bearing in mind that the results also concern the Kinesis parameter of the Transitivity Hypothesis. In our analysis, we will pay special attention to a frequent collocate of *sit*, namely *down*, which marks the clause aspectually telic irrespective of the tense of the clause or grammatical aspect (cf. Hopper 1985). If we find that *down* is particularly often used with *sit* in the simple strategy, this would suggest that transitive syntax may indeed have been supported by a high degree of semantic transitivity in case of *sit*.⁵

We have categorized our data according to the analysis of aspectual categories in Rothstein (2004). First, we take an action to be a description of an aspectually non-stative event. In case of *sit*, this means that descriptions of change in posture are considered to be higher in transitivity in terms of kinesis (e.g. *he sat down*) than descriptions of stative posture (e.g. *he was sitting in a chair*). Second, we

5. We are not sure if it is entirely appropriate to analyse *sitting down* (to indicate a change in posture) as aspectually punctual in the context of the Transitivity Hypothesis. According to textbook definitions, punctual actions have no internal temporal structure or discernible stages (e.g. Comrie 1976: 42–43), and it seems to us that the act of *sitting down* does have temporal duration, albeit a very short one. However, the same can be said for *kick*, which is given as an example of a punctual action in Hopper and Thompson (1980: 252). If *sitting down* is considered to be aspectually punctual, Parameter D (Punctuality) will also be relevant to our analysis, so that clauses which describe a change in posture (e.g. *I sat me down*) are higher in semantic transitivity than those describing stative posture (e.g. *I am sitting by the fire*).

define an aspectually telic clause as one that includes an inherent endpoint (telic point) to the eventuality, which can be either potential or already realized (see also Dahl 1981). For instance, both *he sat down* and *he is sitting down* are aspectually telic. By contrast, *he is sitting in council* or *we sat by the fire* would be aspectually atelic. Finally, in our discussion of the recent transitivization of *sit*, we pay particular attention to the degree of potency of the subject referent (Parameter H: Agency) and the affectedness of the object (Parameter I: Affectedness of O). We will argue that one reason for why *sit* developed transitive uses in clauses like *he sat me down* was the fact that the transitive verb *set* was associated with descriptions of events with highly potent agents, who were in direct physical control of the patient and affected a change of state in it. *Sit* was therefore better suited for descriptions of indirect causation, as in cases where an agent directs someone to sit down instead of physically making them to do so, and where the patient can act on their own free will.

3. Corpora and databases

Our study tracks the development of *sit* from 1500 to 2009. Because of the long time span, there are some obvious data-related challenges that need to be taken into account in the investigation. First, there is currently no corpus that would cover such a long period of time. Second, there is no corpus of British English that would extend all the way from the Early Modern period (or the start of the Late Modern period) to the present day.⁶ Consequently, we have had to use corpora of both British and American English in our research. As these corpora have been compiled according to different principles, and they also represent different linguistic registers, we will study the data in smaller batches and discuss some of the developments on a rather general level: due to the differences in corpus design, it would be highly problematic to compare frequencies across the corpora.

Our Early Modern data (1500–1700) are taken from the *EEBO Corpus*. The EEBO Corpus is based on the Text Creation Partnership version of the *Early English Books Online* (EEBO-TCP; Phase 1 Collection), and it is available for research purposes via the BYU corpus interface. The corpus includes circa 755 million words of text, and each text includes metadata with which the author and the publication year of the text can be identified. The corpus is also annotated for parts of speech, but in this study we did not exploit this feature. The great benefit of the EEBO Corpus is

6. The *Corpus of Late Modern English Texts* (Diller, De Smet & Tyrkkö 2011) does cover the period from 1710 to 1920. However, even though the corpus is relatively large (it includes over 34 million words of text), it did not yield relevant results.

its big size. To compare, the Early Modern part of the Helsinki Corpus, which was used, for instance, in Peitsara (1997), only includes 551,000 words of text (see Kytö 1996). On the downside, the EEBO corpus is rather unstructured when compared to a carefully compiled resource like the Helsinki Corpus. For instance, no attempt has been made to balance the EEBO Corpus according to genre. Furthermore, the corpus may include several editions of a single text, which means that if the search results are not checked and sorted manually, the corpus may yield too conservative results. In this study, we have removed all duplicates to ensure that the data are not skewed because of multiple editions.

Our 18th-century data come from the *Old Bailey Corpus* (OBC; Huber et al. 2012). The OBC includes accounts of spoken witness testimonies as they were reported in a periodical called *The Proceedings of the Old Bailey*. The 18th-century section of the OBC (starting from 1720) includes circa 5.3 million words of text. The corpus is a particularly interesting resource to study because it provides researchers with representations of spoken language from various layers of society and brings us as close to the speech of 18th-century Londoners as possible (see e.g. Huber 2007 for a detailed discussion). In our study, we are mainly interested in seeing whether the OBC includes examples of the emerging transitive use of *sit*. Although the corpus comes with rich sociolinguistic annotation, the topic of our inquiry is unfortunately too infrequently attested to permit a sociolinguistic analysis of the spreading innovation.

Finally, our data for 1810–2009 come from *The Corpus of Historical American English* (COHA; Davies 2010). COHA includes circa 400 million words of text, and it presents a relatively balanced sampling of texts from four genres: fiction, newspapers, non-fiction books and popular magazines. However, the relative proportion of each genre does not remain entirely stable in the 19th-century sub-corpus; for instance, newspapers are only included in the corpus from the 1860s onwards. Like the EEBO Corpus, COHA is also POS-annotated, and the metadata include information about the text's publication year and author. In this study, we examined COHA in order to investigate the most recent developments concerning the transitivization of *sit*.

All corpora were queried by using lexeme-based wildcard searches. The first part of the query included the forms *sit** and *sat**, which ensured that we would retrieve at least the great majority of all grammatical forms and spelling variants of *sit* and *sat* (e.g. *sit*, *sitteth*, *sittes*, *sitting*, *sat*, *sate*, *satte*).⁷ The second part of the query included the first letter of the reflexive forms that could follow *sit*: *m**, *y**,

7. An anonymous reviewer points out that dialectal forms *sutten*, *sot*, and *seet* were still in use in the Early Modern Period. We rechecked the EEBO Corpus for these forms but could not find relevant tokens.

t*, h*, u*, o*.⁸ It was necessary to use the wildcard function in these cases as well, as the texts from the Early Modern period include a great degree of spelling variation. We also studied the use of the transitive verb *set* in the EEBO Corpus. As the frequency of *set* is extremely high, we decided to focus on instances where *set* is followed by an object pronoun and *down*. We used the following query to extract the relevant forms from the corpus: set* m*| y*|t*|h*|u* do*. Once all the relevant forms were retrieved from the corpora, we normalized the frequency data according to twenty-year periods. We decided to use this periodization simply because of ease of visualization – it holds no other significance.

4. Simple and SELF-reflexives with *sit* and the transitivity of *sit down*

4.1 *Sit* in the Early Modern period: Simple and SELF-strategies in the EEBO Corpus

Previous literature has mostly discussed the variation between the simple and the SELF-strategies in terms of systematic replacement: the SELF-strategy gradually took over the function of reflexive marking from the simple strategy in all contexts of use. However, this macro-trend does not accurately depict the changes in the use of *sit* in the Early Modern period. Figure 1 shows that the introduction of the SELF-strategy took place relatively late in case of *sit*: while most verbs had already

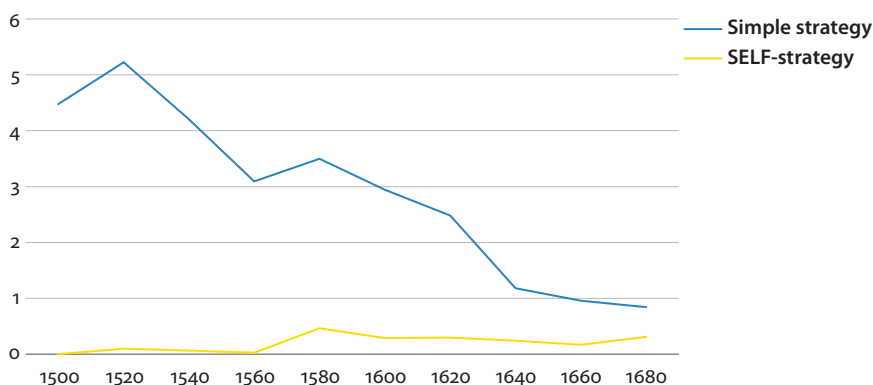


Figure 1. The simple and SELF-strategies with *sit*, 1500–1700. EEBO Corpus. Normalized to 1/1,000,000 words

8. The queries were designed to retrieve the following forms and all their spelling variants: *me/ myself, you/yourself/yourselfes, thee/thyself/thyselfes, him/her/himself/herself, us, ourselves, them/ themselves/theirselfes*.

started to favour the SELF-strategy over the simple strategy by the early 16th century, with *sit* the SELF-strategy remains relatively marginal for the entire Early Modern period. The frequency of *self*-forms does increase in the late 16th century and remains relatively stable throughout the 17th century, but it never surpasses the frequency of the simple strategy (see Appendix 1 for absolute numbers).

The simple strategy, on the other hand, gradually decreases in frequency. This development starts in the 1520s, and by 1700, the normalized frequency has gone down from 5.2 to 0.9 tokens per one million words. It is unlikely that the starting point of this decreasing trend is a coincidence. As discussed, *self*-forms became generally more frequent than simple pronouns in reflexive patterns in the late 1400s and the early 1500s. It is therefore conceivable that the fall of the simple strategy with *sit* is related to this macro-level process: as the simple reflexives became less frequent with other verbs, they also became less frequent with *sit*. Figure 1 also shows that the demise of the simple strategy follows a reversed S-curve with a rapid onset from c. 1520 to 1560, a plateauing period from c. 1560 to 1600, and again a faster decline thereafter. That we are able to see an S-curve in the first place might indicate that *sit* developed rather independently of other verbs or verb types. In the very least, it follows a different S-curve than most verbs which had adopted the SELF-strategy by the early 1500s (for discussions on S-curves and language change, see e.g. Croft 2000: 183; Denison 2003; Nevalainen 2000: 339).

We can interpret the data depicted in Figure 1 as showing two simultaneous large-scale developments. On the one hand, the simple reflexive strategy is on its way out of English grammar, and this macro-level change also affects the use of *sit*. On the other hand, we see evidence of another trend whereby many verbs that could pattern with reflexive marking in older forms of English gradually lost this potential. It should be noted in this context that, as with many other verbs, reflexive marking with *sit* has always been optional. *Sit* has been used intransitively since the Old English period, and according to Peitsara (1997: 322–323) and data from the EEBO Corpus,⁹ the intransitive pattern has always been the dominant pattern. Consequently, the decline of reflexive uses of *sit* in Figure 1 should be taken to depict a gradual loss of a secondary usage pattern, and implicitly, the reinforcement of the primary intransitive pattern.

Importantly, and somewhat surprisingly, Figure 1 also suggests that the two reflexive strategies are not linguistic variants: the simple strategy is not replaced by the SELF-strategy in the period studied. This raises the question of whether the two strategies might in part have been used to describe different kinds of events in

9. Unfiltered corpus data suggest that only around 10% of all tokens of *sit* were reflexive in the Early Modern period.

Early Modern English, and this is in fact what our data suggest. Examples (9)–(14) illustrate the typical usage of both the simple and the SELF-strategies with *sit*.

- (9) Here will I *sit me downe* and fixe mine eye upon the ruines of you wretched towne. (EEBO: 1594)
- (10) I *sate me downe* there vpon the stayres or steps. (EEBO: 1623)
- (11) We *sat us down* under the shadow of a great tree that stood by it self. (EEBO: 1653)
- (12) ... by the counsaile and aduise off certeine persons he returned and *sat him self downe* in the pastors place with the seniors. (EEBO: 1574)
- (13) With that they *sate themselves* to eat, which Clarinda did very heartily. (EEBO: 1692)
- (14) He *sitteth himself* in council dayly, and disposeth affairs of most weight in his own person. (EEBO: 1655)

In Examples (9)–(11), we see the simple strategy used first in the present tense (as in (9)), and then in the past tense (as in (10) and (11)). In all these cases, the pattern is followed by the adverb *down*, indicating a change in posture and the completion of the action. In (12)–(14), on the other hand, we see the SELF-strategy used in the past tense (as in (12) and (13)) and the present tense (as in (14)). Similarly to Examples (9)–(11), Example (12) is also followed by *down*. Example (13), by contrast, is used without *down*, while Example (14) illustrates another meaning of *sit*: ‘to sit in council’.¹⁰

In order to tease out the subtle differences in the use of the two reflexive strategies with *sit*, we will now turn our attention to the question of what kinds of events were described with the simple and the SELF-strategies from the perspective of lexical aspect. More specifically, we are interested in determining whether the simple strategy was associated with descriptions of aspectually telic events, and consequently, with meanings related to a high degree of semantic transitivity.

10. An anonymous reviewer points out that the *self*-form in examples like (14) can also be read to indicate emphasis. We agree, and we acknowledge that while we have done our best to exclude clearly emphatic uses from our data, some of the forms included in the analysis remain ambiguous.

4.2 Expressing telic and atelic aspect with the simple and the SELF-strategies

In Section 2, we hypothesized that the prolonged use of the simple strategy with *sit* may have been supported by meanings associated with a high degree of semantic transitivity. Our data provide support to this hypothesis, showing that the simple pattern is more strongly associated with telic descriptions than the SELF-strategy. Figure 2 shows that the proportion of telic uses ranges from 93% to 97% in the simple strategy,¹¹ while the corresponding proportion for the SELF-strategy varies from 33% (1600–1620) to 91% (1660–1680). (See Appendix 2 for absolute numbers).

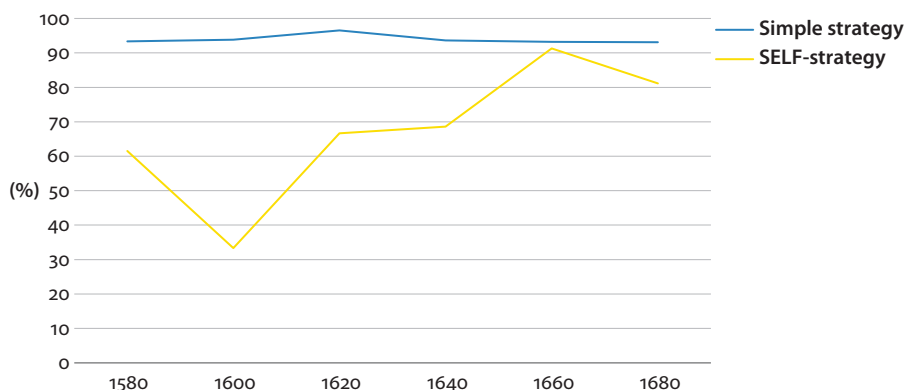


Figure 2. *Sit*: The proportion of telic descriptions with the two reflexive strategies, 1580–1700. EEBO Corpus

In addition to the difference in the use of the two reflexive strategies depicted in Figure 2, the strategies also differ in terms of their co-occurrence with *down*, which functions as an overt marker of telicity with *sit*. While *down* is a frequent collocate of *sit* in both reflexive strategies, our data show that it was particularly associated with the simple strategy in the period studied (1580–1700): 981 out of 1,092 tokens of *sit* are used together with *down* in the simple strategy (89.8%), while in the SELF-strategy, only 82 out of 163 tokens of *sit* co-occur with *down* (50.3%). In other words, instead of being used to express physical posture, the simple strategy was specifically used to indicate a change in posture, and this was overtly encoded in the clause by using the telic marker *down*. Figure 3 shows the proportion of the use of *down* in both strategies. Absolute numbers are given in Appendix 3.

11. Here, we only discuss data from 1580 to 1700 as the frequency of *self*-reflexives is too low to be compared with the simple strategy prior to the 1580s.

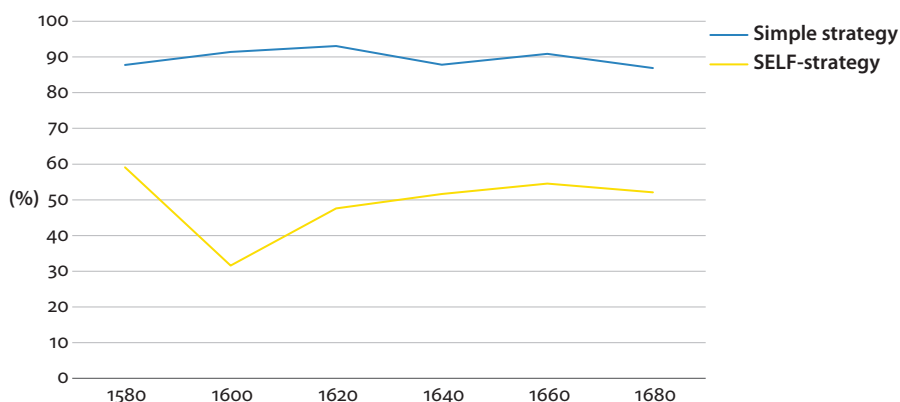


Figure 3. *Sit: Sit down* with the simple and SELF-strategies, 1580–1700. EEBO Corpus

To illustrate the difference in the use of the two strategies, let us take a look at some representative examples of the use of the SELF-strategy in the EEBO Corpus.

- (15) When Iethro saw Moses his sonne in law, *sitting himselfe alone* and iudging the people, from morning vnto even, he did not lesse then reprooue him for it. (EEBO: 1599)
- (16) For it was the manner in those daies the king should determine causes of great waight *sitting himselfe in open court*. (EEBO: 1601)
- (17) After the end of this speech I deliuered Sir Thomas Greshams letters, when as he tooke me by the hand, and led me downe a long court to a palace where there ranne a faire fountaine of water, and there *sitting himselfe in a chaire*, he commanded me to sit downe in another, and there called for such simple musicians as he had. (EEBO: 1600)
- (18) He *sate himself down* in a book-seller's shop. (EEBO: 1601)

In (15), the SELF-strategy is used to express posture, and the event described is aspectually stative (and therefore atelic; see Rothstein 2004: 14–17). In (16), on the other hand, we find the *self*-form used in the sense ‘to sit in court’, while in (17) the SELF-strategy is used to indicate a change in posture without *down*, and in (18) with *down*. Significantly, none of the uses exemplified from (15) to (17) are commonly expressed with the simple strategy, which – as we have seen – is very strongly associated with the kind of usage in (18), that is, a change in posture and the use of *down*. The stative meaning in (15) is very rarely expressed with the simple strategy in our data, and we found no examples of ‘to sit in court’ expressed with the simple strategy.¹²

12. An anonymous reviewer raises the possibility that the question may also be about the fact that simple pronouns cannot be used in an emphatic function. As noted previously, the *self*-forms are indeed often ambiguous between emphatic and non-emphatic readings, and we cannot completely rule out this possibility. We will discuss this issue in more detail in Section 5.

In sum, we conclude that in the Early Modern period the simple strategy was typically used to express a change in posture. In other words, it was particularly often used in descriptions of aspectually telic events with a highly potent agent. The connection between telicity and the simple strategy is particularly evident in the frequent use of the adverb *down*. The SELF-strategy, by contrast, shows greater variation in its use. In addition to descriptions of telic eventualities, the SELF-strategy was also used to describe different kinds of atelic/stative meanings, such as sitting in a chair/throne or sitting in judgment. Furthermore, the frequency of *down* is generally much lower in the SELF-strategy than in the simple strategy. Based on these findings, we conclude that the simple strategy was indeed strongly connected with meanings that are high in transitivity in the sense of Hopper and Thompson (1980), and this may have affected its longevity in the Early Modern period.

4.3 The transitivization of *sit down*

4.3.1 *Overlap between sit and set and early examples of the transitivization of sit down*

In this section, we focus on the second research question of our study, the transitivization of *sit down*. Before discussing the data, however, it is necessary to pay attention to a phenomenon that will be relevant to our analysis: the overlap in the use of the intransitive verb *sit* and the transitive verb *set* that is evident in both Middle English and Early Modern English texts (see OED s.v. *set* 3). We have not been able to identify any functional motivation for this overlap in our data or in previous literature (e.g. Visser 1963: 322). Indeed, while in principle *setting one/oneself down* might indicate a difference in meaning from *sitting one/oneself down*, in that *setting one/oneself down* may not indicate a sitting posture, there are corpus examples like (19) and (20) in which a sitting posture is clearly indicated with *set*.¹³

(19) He settled himselfe to heare the rest of the seruice, and *setting him downe in his chaire*, hee came to his first admiration... (EEBO: 1612)

(20) ... and coming into the church, *he set him down in a seat* just before master clark. (EEBO, 1660)

Our data on *set down* consist of 139 tokens from 1553 to 1685. The data show that when *set* was used in its causative sense (i.e. when it was not confused with *sit*), the clause describes an event that was high in semantic transitivity according to a

13. The overlap may simply have been due to the vowel quality in *sit* and *set*, which may have been similar enough for some speakers to use the two verbs interchangeably. Indeed, our data also include examples of *sit* used instead of *set*, as in “to prepare and sit [pro *set*] themselves for the exhibition” and “to sit [pro *set*] himself with horse and arms”.

number of parameters: a potent and volitional agent (Parameters E and H) affects a change of state in a highly individuated patient (Parameters I and J). Furthermore, the clause is a description of an action (Parameter B) that includes an inherent endpoint (Parameter C). If the patient was human, this meant that they were either carried, physically assisted or forcefully set down, as exemplified in (21)–(23).

(21) ... and *the quene of Orqueneey* toke Arthur by the hande & *set him downe* by her. (EEBO: 1560)

(22) *He which carried the valiant Frenchman prisoner*, was constrained to *set him downe* vpon the hard ground. (EEBO: 1601)

(23) I raised him vp, then *set him downe* againe, then puld him here. (EEBO: 1606)

Bearing in mind that *set down* was used to describe events that were high in semantic transitivity in the EEBO data, let us now observe a new meaning for *set down* that emerged in the late 17th century: ‘to cause or allow (someone) to alight from a vehicle’ (OED, s.v. *set down* (4b)). Examples (24) and (25) illustrate this new usage.

(24) ... then I desired the coachman *to set me down*. (EEBO: 1664)

(25) *You shall set me down* in Lincolus-inn-fields then. (EEBO: 1672)

The meaning expressed in (21)–(23), where the patient was physically assisted or forcefully set down, contrasts in two important ways with the meaning illustrated in (24) and (25). First, in (24) and (25) there is no direct contact between the coachman and the passenger, at least not necessarily; the agent is therefore not described to be as potent in these examples as in (21)–(23). Second, although the coachman is in control of the coach, the passenger is in full control of alighting from the coach, and so the patient is not affected by the action to the same extent as in Examples (21)–(23). In short, *set* had previously not been associated with the kind of indirect causation exemplified in (24) and (25).

We would like to argue that this usage provided a particularly favourable context for *sit* to be used transitively instead of *set*. The occasional confusion in the use of *sit* and *set* can be regarded as a precondition for the early transitive uses of *sit*, but we suggest that what truly facilitated the transitivity process was the fact that *sit* was not associated with the same semantic constraints regarding the high potency of the agent and the affectedness of the patient as *set*. The first examples of transitive uses of *sit* in our data come from the 1720s and the 1730s (from the Old Bailey Corpus).

(26) ... *when he sat her down*, he drove away with the cheeses. (OBC: 1727)

(27) ... and then from thence we all, but Howard, went home in a Coach; *they sat me down at my Door*, and then they discharged the Coach. (OBC: 1733)

It should be noted that this usage is very rare in our data with only four tokens in the 18th-century sub-corpus (there are three more tokens in the 19th-century section of the OBC). Nevertheless, the data do show that the earliest transitive uses of *sit* took place in a context that was associated with a low degree of semantic transitivity. After *sit* became associated with transitive uses in these contexts, its usage started to extend to contexts illustrated in (21)–(23) – the ones where the agent is in direct contact with the patient and physically assists them to sit down, as in (28) and (29). The difference between Examples (21)–(23) and Examples (28) and (29) is, of course, that in (28) and (29) the patient's sitting posture is implied. In (22) and (23), by contrast, it is also possible that the patient was set down in a lying posture. There are altogether ten tokens of this particular usage in the entire OBC, five of which are from the 18th century.

- (28) They [...] carried me into the Room where the Boy was, and *sat me down by the Fire*. (OBC: 1735)
- (29) When they had got me into the house, *they sat me down* gently into a chair, they did not hurt me at all. (OBC: 1745)

To sum up this section, we have seen that *sit down* started to be used in transitive constructions in the early 1700s. We have suggested that this development was facilitated by two factors: (i) the overlapping use of *sit* and *set*, and ii) the creation of a new meaning ('to allow someone to alight from a vehicle') with which *set* was not perfectly compatible. Furthermore, it is possible, in principle at least, that the transitivization of *sit* may also have been facilitated by its persisting association with transitive syntax in the simple reflexive strategy. In the following section, we will discuss the most recent developments of this transitivization process by studying data from 19th- and 20th-century American English.

4.3.2 *Recent developments in the transitivization of sit down: Evidence from COHA*

The data from COHA show that *sit down* was used in 19th-century American English in much the same way as it was in 16th- and 17th-century British English. First, both the simple and the SELF-strategies are attested in the data, although they are on a path of steady decline (Figure 4; see Appendix 4 for absolute frequencies).

When it comes to transitive uses of *sit down*, COHA includes examples that are similar to Examples (28) and (29), which were taken from the OBC. In (30)–(32), the agent either carries the patient or lifts them up from the ground and then makes them sit down.¹⁴

14. Examples (30) and (32) could also be analysed in terms of the overlapping use of *sit* and *set*. Example (31), on the other hand, clearly involves setting the patient in a sitting posture.

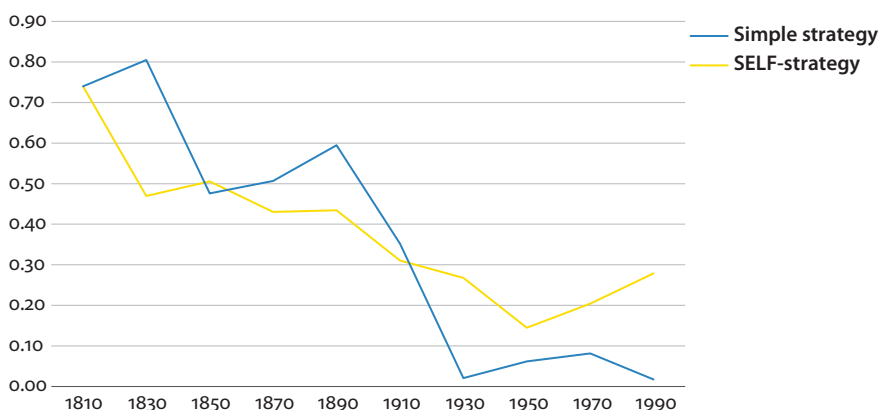


Figure 4. *Sit down*: The simple and SELF-strategies in American English, 1810–2009. COHA. Normalized to 1/1,000,000 words

- (30) ... and scrambling up the rocks where the waves could not reach her, *he sat her down* and took breath. (COHA: Fiction, 1847)
- (31) ... *the old lady* would shake her, and bring her back, and *sit her down* on the chair so hard as to make her cry with pain... (COHA: Fiction, 1853)
- (32) *He* lifted him high and *sat him down* emphatically, but not injuriously. (COHA: Fiction, 1864)

Starting from the 1870s, we begin to find examples of transitive uses of *sit down* where the agent is not necessarily as high in potency as in Examples (30)–(32). While it is possible that the patient was physically made to sit down in Examples (33) and (34), it is equally possible that they were simply gestured or told to sit down.

- (33) And when *she* saw his face pale with suffering, she forgot all about the rehearsal, and shook his hand with sisterly heartiness – the word “sisterly” came to her mind most opportunely – and looked at him with the utmost gladness, and *sat him down* by the window. (COHA: Fiction, 1873)
- (34) *The Governor* took me at once to his own room, and *sat me down* at the table. (COHA: Fiction, 1899)

Despite the occasional examples of (potentially) indirect causation represented in (33) and (34), most transitive uses of *sit down* still express direct causation in the 1900s and the 1910s. However, things start to change in the 1920s and the 1930s. First, we begin to see ambiguous cases where the patient is first *led*, *taken*, *brought*, or *escorted* into a room and then *sat down*. All these verbs can be conceptualized as describing an event where the agent is in direct contact with the patient while leading them into a room, and then physically assists them to take a seat (i.e. the agent is highly potent), but they can also be read so that the patient is following

the agent and verbally directed or motioned to sit down (i.e. the potency of the agent is lower). Examples (35)–(37) illustrate such ambiguous uses from the 1920s and the 1930s, and Figure 5 presents our data from COHA divided according to the frequency of direct causatives, indirect causatives and ambiguous cases (see Appendix 5 for absolute frequencies).¹⁵

- (35) *Daniel led* Aunt Matty back into the house and *sat her down* on a sofa in the library. (COHA: Fiction, 1930)
- (36) I reckon what they said was true, but *he brought* that child home and *sat her down* at the table with all the rest of them. (COHA: Fiction, 1935)
- (37) “Dry your eyes,” *Myrtle said, leading* Margaret to the divan, *sitting her down*, and then seating herself beside her friend. (COHA: Fiction, 1938)

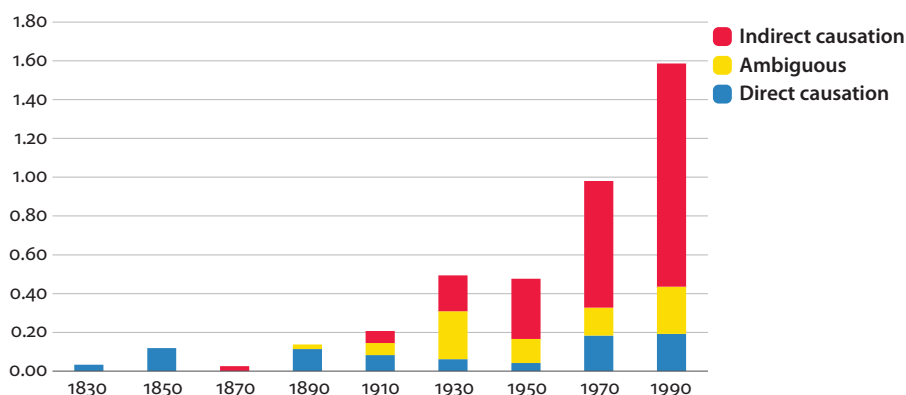


Figure 5. Transitive *sit down*: Direct causatives, indirect causatives and ambiguous cases, 1830–2009. COHA. Normalized to 1/1,000,000 words

It is, of course, true that the transitive construction itself assigns a certain degree of control to the agent, and it is consequently easy to find examples where the patient is given little choice in deciding whether to sit down or not; in (38), for example, the patient is practically forced to sit down for a discussion. However, the patient’s act of assuming a sitting posture is not overtly described to be physically affected by the agent of the clause, and hence we have categorized this kind of usage as an example of indirect causation.

- (38) *His father sat him down* and made him choose which parent he would live with. (COHA, Magazines, 2005)

15. In Figure 5, we have only classified cases like (35)–(37), where *sit N down* is preceded by a verb like *lead* or *bring*, as ambiguous.

Indeed, in many cases the potency of the agent is difficult to assess; it is always possible that the agent physically directs the patient to take a seat, but this cannot be deduced from the context. However, our data do include examples where the agent is clearly very low in potency, as in (39), where two journalists discuss the possibility of inviting Bill Clinton, who at the time was a presidential candidate for the Democratic Party in the U.S., for an interview. Such examples can be classified as indirect causatives with high confidence.

- (39) Ken, do you think *we can sit Clinton down* and have him talk to the camera?
(COHA: Magazines, 1992)

To summarize, our data suggest that the recent transitivization of *sit down* has proceeded from infrequent expressions of direct causation to relatively frequent expressions of indirect causation. Data from the 1920s to the 1940s suggest that this process may have been facilitated by ambiguous uses where the agent was described as leading or escorting the patient into a room, but physical contact between the two participants was not made explicit. The most recent dataset from 1990–2009, on the other hand, shows that in Present-day American English, the main function of the transitive *sit down* is to express indirect causation. According to our analysis, we conclude that *sit down* has thus filled a functional slot that *set* was unable to fill due to its association with expressions of high semantic transitivity.

5. Discussion and conclusions

In this chapter we have discussed the development of a single verb in the history of English: *sit*. Our goal has been to study the variation between the simple and the SELF reflexive strategies in the Early Modern period and the transitivization of *sit down* in Late Modern and Present-day English. In our analysis of the two reflexive strategies, we paid particular attention to the linguistic contexts in which the strategies were used. The motivation for studying this question arose in part from the fact that earlier studies of variation in reflexive strategies were based on relatively small datasets (Peitsara 1997 in particular). Consequently, while previous studies were able to provide fascinating insights into many aspects of the variation and change of the English reflexive system, they necessarily suffered from the limited amount of data on which the analyses were based. By focusing on a single verb, and by comparing the variation between the simple and the SELF-strategies in different linguistic contexts, we hope to have shown that it is worthwhile to look back on previous research with new data and see if we can uncover more detailed information about a historical phenomenon that has already been subject to meticulous research in the past.

In Section 4.1, we showed that, as far as *sit* is concerned, the simple and the SELF-strategies did not occupy the same functional space and that they were not truly linguistic variants. In particular, the simple strategy was strongly associated with the use of *down*, which served as an aspectual marker to indicate that the action included an inherent endpoint (a change in posture). However, there is one confounding factor that we briefly mentioned in our discussion of the use of the two reflexive strategies; namely, the fact that the *self*-forms might in some cases indicate emphasis instead of reflexivity. More specifically, we cannot entirely rule out the possibility that the *self*-forms that were used in some of the atelic descriptions ('to sit in council/judgment') are emphatic instead of reflexive, and the fact that simple reflexives are not attested in this particular usage in our data may suggest that the *self*-forms should indeed be understood to indicate emphasis. However, even if this is the case, it will not have a substantial effect on our results: if we remove these data from our analysis, 79% (128/163) of all *self*-reflexives are used in telic descriptions in 1580–1700, while the corresponding proportion for simple reflexives is 94% (1,027/1,092). This difference is still statistically highly significant ($p < 0.0001$).

As argued, our findings can be explained particularly well by the Transitivity Hypothesis: a high degree of semantic transitivity supported the persistence of transitive syntax in the simple reflexive strategy. We are not aware of other studies that explain historical variation and change in the English reflexive system from the perspective of the Transitivity Hypothesis, and our results suggest that it would be interesting to investigate other verbs and verb types from this angle as well. While it is possible that *sit* has developed along a path of its own – as might be suggested by the reversed S-curve in Figure 1 – it is also possible that the Transitivity Hypothesis could explain the variation in the use of the reflexive strategies with other verbs. One obvious candidate to study would be *lie*, which is in many ways similar to *sit*: not only can it be used with the simple strategy in PDE (albeit in a very restricted way), it has also been confused with its transitive counterpart *lay* since the Middle English period (just like *sit* and *set*; see e.g. Visser 1963: 322).

However, it should also be pointed out that although the Transitivity Hypothesis has been credited for establishing significant connections between semantic event structure and syntactic structure, the theory has also faced some criticism. For example, as already pointed out in Tsunoda (1985), Hopper and Thompson (1980) do not consider the possibility that the transitivity parameters might affect syntactic structure to different degrees. Furthermore, the Transitivity Hypothesis makes some counterintuitive predictions, suggesting, for example, that a one-actant clause like *Susan left* is in fact higher in semantic transitivity than a two-actant clause like *John likes beer* (see e.g. Lazard 2002: 178). It is obvious that if the Transitivity Hypothesis is found to be useful in the description of historical transitivization and/

or intransitivization processes, there are still open questions in the theory that need to be resolved (see e.g. LaPolla, Kratochvíl & Coupe 2011).

In our second case study, we argued that the transitivization of *sit* was facilitated by the fact that the transitive verb *set* was associated with features of high transitivity and was therefore ill-suited for expressing the meanings which emerged with the transitive *sit down*, that is, different kinds of indirect causation. In addition to semantic transitivity, there have been large-scale processes at work in English grammar which may also have had an indirect impact on the transitivization of *sit*. First, McMillion (2006) shows that the number of labile verbs (verbs that can be used both transitively and intransitively; for instance, *he broke the window* vs. *the window broke*) has steadily increased in English throughout the Modern period, and it is plausible that this tendency may also have supported the transitivization of *sit*. Second, as observed in Algeo (2006), for example, many verbs that used to take prepositional complements in earlier forms of English can now be used transitively, particularly in American English (e.g. *he graduated high school* vs. *he graduated from high school*). Some studies (e.g. Kirchner 1955) suggest that this trend can be traced all the way back to the Early Modern period (see also Callies 2018; Rohdenburg 2009). Indeed, in addition to the semantic factors affecting the transitivization of *sit* that were discussed in this chapter, the potential effect of these macro-processes should certainly be acknowledged, even if the extent of their influence may not be possible to assess or quantify with precision.

References

Corpora

- COHA = Davies, Mark. 2010–. *The Corpus of Historical American English: 400 Million Words, 1810–2009*. <<http://corpus.byu.edu/coha/>> (30 May 2018).
- EEBO corpus = *Early English Books Online Corpus*. <<https://corpus.byu.edu/eebo/>> (30 March 2018).
- Huber, Magnus, Nissel, Magnus, Maiwald, Patrick & Widlitzki, Bianca. 2012. *The Old Bailey Corpus. Spoken English in the 18th and 19th centuries*. <www.uni-giessen.de/oldbaileycorpus> (30 March 2018).

Other references

- Algeo, John. 2006. *British or American English? A Handbook of Word and Grammar Patterns*. Cambridge: CUP.
- Ariel, Mira. 2012. Research paradigms in pragmatics. In *The Cambridge Handbook of Pragmatics*, Keith Allan & Kasia M. Jaszczolt (eds), 23–46. Cambridge: CUP.

- Callies, Marcus. 2018. Patterns of direct transitivization and differences between British and American English. In *Changing Structures. Studies in Constructions and Complementation*, [Studies in Language Companion Series 195], Mark Kaunisto, Mikko Höglund & Paul Rickman (eds), 151–167. Amsterdam: John Benjamins.
- Comrie, Bernard. 1976. *Aspect*. Cambridge: CUP.
- Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. London: Longman.
- Crystal, David. 2008. *A Dictionary of Linguistics and Phonetics*, 6th edn. Oxford: Blackwell.
- Dahl, Östen. 1981. On the definition of the telic-atelic (bounded-nonbounded) distinction in tense and aspect. In *Syntax and Semantics 14: Tense and Aspect*, Philip Tedeschi & Annie Zaenen (eds), 79–90. New York, NY: Academic Press.
- Denison, David. 2003. Log(ist)ic and simplistic S-curves. In *Motives for Language Change*, Raymond Hickey (ed.), 54–70. Cambridge: CUP.
<https://doi.org/10.1017/CBO9780511486937.005>
- Diller, Hans-Jürgen, De Smet, Hendrik & Tyrkkö, Jukka. 2011. A European database of descriptors of English electronic texts. *The European English Messenger* 19: 21–35.
- Hopper, Paul J. 1985. Causes and affects. In *Causatives and Agentivity. Papers of the Parasession on Agency and Causativity of the 21st Annual Meeting of the Chicago Linguistics Society*, William H. Eilfort, Paul D. Kroebeer & Karen L. Peterson (eds), 67–88. Chicago, IL: Chicago Linguistics Society.
- Hopper, Paul J. & Thompson, Sandra A. 1980. Transitivity in grammar and discourse. *Language* 56(2): 251–299.
- Huber, Judith. 2017. *Motion and the English Verb: A Diachronic Study*. Oxford: OUP.
- Huber, Magnus. 2007. The Old Bailey Proceedings, 1674–1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English. In *Annotating Variation and Change* [Studies in Variation Contacts and Change in English 1], Anneli Meurman-Solin & Arja Nurmi (eds). Helsinki: Varieng. <<http://www.helsinki.fi/varieng/series/volumes/01/huber/>> (31 May 2019).
- Ito, Eiko. 1978. Reflexive verbs in Chaucer. *Studies in English Literature by the English Literature Society of Japan: The English Number*, 65–89. Tokyo: Kenkyusha.
- Keenan, Edward. 2002. Explaining the creation of reflexive pronouns in English. In *Studies in the History of the English Language: A Millennial Perspective*, Donka Minkova & Robert P. Stockwell (eds), 325–354. Berlin: De Gruyter.
- Kirchner, Gustav. 1955. Direct transitivation. *English Studies* 36: 15–23.
<https://doi.org/10.1080/00138385508596940>
- Kytö, Merja. 1996. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*, 3rd edn. Helsinki: Department of English, University of Helsinki.
- LaPolla, Randy J., Kratochvíl, František & Coupe, Alexander R. 2011. On transitivity. *Studies in Language* 35(3): 469–491. <https://doi.org/10.1075/sl.35.3.0oint>
- Lazard, Gilbert. 2002. Transitivity revisited as a more strict approach in typological research. *Folia Linguistica* 36(3–4): 141–190. <https://doi.org/10.1515/flin.2002.36.3-4.141>
- Lyons, John. 1968. *Introduction to Theoretical Linguistics*. Cambridge: CUP.
- Mitchell, Bruce. 1985. *Old English Syntax* 1. Oxford: Clarendon.
- McMillion, Alan. 2006. *Labile Verbs in English: Their Meaning, Behavior and Structure*. Stockholm: Engelska institutionen.
- Möhlhig-Falke, Ruth. 2012. *The Early English Impersonal Construction: An Analysis of Verbal and Constructional Meaning*. Oxford: OUP.

- Mondorf, Britta. 2016. "Snake legs it to freedom": Dummy *it* as pseudo-object. *Corpus Linguistics and Linguistic Theory* 12(1): 73–102.
- Mustanoja, Tauno F. 1960. *A Middle English Syntax, Part I: Parts of Speech*. Helsinki: Société Néophilologique.
- Næss, Åshild. 2007. *Prototypical Transitivity* [Typological Studies in Language 72]. Amsterdam: John Benjamins.
- Nevalainen, Terttu. 2000. Processes of supralocalisation and the rise of Standard English in the Early Modern period. In *Generative Theory and Corpus Studies: A Dialogue from 10 ICEHL* [Topics in English Linguistics 31], Ricardo Bermúdez-Otero, David Denison, Richard M. Hogg & C. B. McCully (eds), 329–372. Berlin: Mouton de Gruyter.
- Nordlinger, Rachel. 2011. Transitivity in Murrinh-Patha. *Studies in Language* 35(3): 702–734. <https://doi.org/10.1075/sl.35.3.o8nor>
- OED = *OED Online*. Oxford: OUP. <<https://www.oed.com/>> (15 March 2019).
- Palander-Collin, Minna. 1999. *Grammaticalization and Social Embedding: I THINK and METHINKS in Middle and Early Modern English* [Mémoires de la Société Néophilologique de Helsinki 55]. Helsinki: Société Néophilologique.
- Peitsara, Kirsti. 1997. The development of reflexive strategies in English. In *Grammaticalization at Work. Studies of Long-term Developments in English*, Matti Rissanen, Merja Kytö & Kirsi Heikkonen (eds), 277–370. Berlin: Mouton de Gruyter.
- Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Reuland, Eric. 2000. The fine structure of grammar: Anaphoric relations. In *Reflexives: Forms and Functions*, Vol. 1 [Typological Studies in Language 40], Zygmunt Frajzyngier & Traci S. Curl (eds), 1–40. Amsterdam: John Benjamins.
- Rohdenburg, Günter. 2009. Nominal complements. In *One Language, Two Grammars? Differences between British and American English*, Günter Rohdenburg & Julia Schlüter (eds), 194–211. Cambridge: CUP.
- Rothstein, Susan. 2004. *Structuring Events: A Study in the Semantics of Lexical Aspect*. Oxford: Blackwell.
- Spies, Heinrich. 1897. *Studien zur Geschichte des englischen Pronomens im xv. und xvi. Jahrhundert* [Studien zur englischen Philologie 1]. Halle a.S.: Max Niemeyer.
- Tsunoda, Tasaku. 1985. Remarks on transitivity. *Journal of Linguistics* 21(2): 385–396. <https://doi.org/10.1017/S0022226700010318>
- Van Gelderen, Elly. 2000. *A History of English Reflexive Pronouns. Person, Self, and Interpretability*. Amsterdam: John Benjamins.
- Visser, Fredericus Theodorus. 1963. *An Historical Syntax of the English Language, Part One: Syntactical Units with One Verb*. Leiden: E. J. Brill.

Appendix 1. The absolute frequencies of the simple and the SELF-strategies, 1500–1700: EEBO Corpus

	1500	1520	1540	1560	1580	1600	1620	1640	1660	1680
Simple	10	53	67	133	196	244	202	173	132	145
Self	0	1	1	1	26	24	24	35	23	53

Appendix 2. The absolute frequencies of telic descriptions in the simple and the SELF-strategies, 1580–1700: EEBO Corpus

	1580	1600	1620	1640	1660	1680
Simple	183	229	195	162	123	135
Self	16	8	16	24	21	43

Appendix 3. Occurrence of *down* in the simple and the SELF-strategies, 1580–1700: EEBO Corpus. Absolute frequencies

	1580	1600	1620	1640	1660	1680
Simple	172	223	188	152	120	126
Self	13	6	10	16	12	25

Appendix 4. The absolute frequencies of the simple and the SELF-strategies, 1810–2009: COHA

	1810	1830	1850	1870	1890	1910	1930	1950	1970	1990
Simple	6	24	16	20	26	17	1	3	4	1
Self	6	14	17	17	19	15	13	7	10	16

Appendix 5. Transitive *sit down*: Direct and indirect causatives and ambiguous cases, 1830–2009: COHA. Absolute frequencies

	1830	1850	1870	1890	1910	1930	1950	1970	1990
Direct causation	1	4	0	5	4	3	2	9	11
Indirect causation	0	0	1	0	3	9	15	32	66
Ambiguous	0	0	0	1	3	12	6	7	14
Total	1	4	1	6	10	24	23	48	91

Index

A

- A Representative Corpus of
Historical English Registers
see ARCHER
- ACRONYM 80
- adolescent 146, 151, 153, 160
- adult 99, 146–147, 153, 160
- adverbial 161–162, 170, 174,
178, 182, 191, 194, 211–212, 216,
259, 268
- adverbial shift 216, 221
- age 143–148, 151–153, 155–156,
160, 166–167, 228–229 *see also*
adolescent, adult
- agency 283–285, 292–297
- animacy 6, 13–18, 22–24
- AntConc 61–62
- APRIL 80–81
- ARCHER 173, 182, 185, 191
- aspect 125, 154, 283–285,
289–292, 298 *see also* stative,
telicity
- attitude 43, 53–54, 61, 76, 143,
162, 172 *see also* stance
- authority 3, 5–6, 14–15, 18–20,
22–25, 122
- AVIATOR 80

B

- bigrams 230, 239–240
- bleaching 161, 171–172,
190, 193, 227–228, 268
see also delexicalization,
desemanticization
- BNC 40, 145–148, 150–157,
159–160, 163, 166–167, 170,
255–257, 259–260
- booster 144, 152, 160, 171, 194,
224, 228, 230, 245 *see also*
maximizer
- British National Corpus *see*
BNC

C

- CEN *see* Corpus of English
- Novels
- Census 114–116, 118, 122–123,
137
- CLMET3.0 33–34, 43, 47–49,
51–53, 262, 267, 271, 285
- close reading 29, 46 *see also*
distant reading
- COCA 19–21, 252–257, 259–
260, 265–266
- COHA 3, 16–18, 20–21,
23–24, 223–224, 229, 231–232,
234–235, 238–239, 241–244,
262–267, 269–271, 278, 286,
294–297, 302
- colloquialization 269–270, 272
- community 4–5, 128–129, 137
- complement-taking predicate
257–259, 261, 268, 270–271
- conceptual maps 34, 39–40,
51–52
- conjunct 212–213, 215–216, 219
see also disjunct
- content analysis 30, 36–38
- Corpus of Contemporary
American English *see* COCA
- Corpus of Early English Medical
Writing 60–61
- Corpus of Early Modern English
Medical Texts *see* EMEMT
- Corpus of English Novels
262, 267
- Corpus of Historical American
English *see* COHA
- Corpus of Late Modern English
Medical Texts *see* LMEMT
- Corpus of Late Modern English
Texts *see* CLMET3.0
- Corpus of News on the Web *see*
NOW
- corpus-driven 30, 35–36 *see*
also data-based, data-driven

- covarying collexeme analysis
240–244
- critical discourse analysis
113, 117–118, 138
- culturonomics 29, 40

D

- data-based 35–36, 53 *see also*
corpus-driven, data-driven
- data-driven 30, 32, 35–36, 39,
43, 53–54 *see also* corpus-
driven, data-based
- decategorialization 190, 193,
268, 271
- delexicalization 161, 193,
227–228, 245
- desemanticization 171, 190,
268, 271
- dictionary approach 36, 39, 53
- discourse marker 143, 157–158,
160–162, 257, 259, 261
- disjunct 199, 212–213, 216, 219
see also conjunct
- distant reading 29, 46 *see also*
close reading
- distributional semantics 18–19,
38–39, 53–54, 235

E

- Early English Books Online *see*
EEBO
- Early Modern English Prose
Selections *see* EMEPS
- ECF 173, 182–187, 263
- Edinburgh Medical Journal *see*
EMJ
- EEBO 29–30, 32–34, 42–44,
48–49, 52–53, 173, 178–180,
182, 185, 193, 201, 203, 262–
267, 271, 285–293, 302
- Eighteenth Century Collections
Online 263, 267

Eighteenth Century Fiction *see*
ECF

EMEMT 58, 60–66, 72, 74–75

EMEPS 201–203, 205, 210–211,
214–215, 217

EMJ 60–61

emphasis 143–144, 153–158,
161–162

epistemic 20, 22, 25, 154–155,
158, 260–261

F

female 148–149, 160, 163,
228–229

G

gender 145, 147–149, 166–167,
229 *see also* female, male

genre 33–34, 49, 60, 74–75,
202, 223–245, 256–257, 267,
269, 282, 286

academic 227, 257, 267, 269

conversation 144–145, 149,
227

drama 251, 267, 271

institutional texts 60–61,
73–75

magazine 257, 267, 296–297

newspaper 80, 115, 133–134,
256–257, 267, 271, 286

novel 173, 200, 267, 277

religious text 179, 201–2,
204–207, 209

Google Books 3–5, 10, 29–30,
267–268

grammaticalization 6, 12,
15, 161–162, 169, 171–172,
188, 190–191, 193–194, 227,
243, 245, 251–252, 258–262,
268–272, 281

H

Hansard 262, 267

Harry Potter adverbs 200,
216–219

HC *see* Helsinki Corpus

Helsinki Corpus 173, 175,
177–178, 200–202, 209, 278,
282, 286

hyperbole 152, 154, 156–157, 163

K

kernel density estimation

29, 36, 39–40, 51, 53–54

keyword 37–39, 44–45, 47–51,
58

L

lexical bundle 59, 63–64, 66, 72

lexical diversity 10, 12, 228,
237–239, 243–244

lexical replacement 224–226,
232, 235, 242

LMEMT 58, 60–63, 69–72,
74–76

M

male 143, 148–149, 163, 229

Mallet 43, 46–49

maximizer 144, 152, 171, 223 *see*
also booster

ME *see* Middle English

MED 174, 207, 210, 215, 271

media content analysis 30,
36–37, 54

metaphor 79, 95, 98, 102, 107,
174

MTiP 201–207, 213, 218–219

MI measure 59, 61–63, 69,
118–119, 121, 125–126, 131

Middle English 69, 169, 172–
178, 180, 188, 190, 200–201,
204–207, 209–210, 212, 218,
225, 251, 271–272, 278, 281,
292, 298

Middle English Dictionary *see*
MED

N

native speaker 134–137, 226 *see*
also non-native speaker

NCF 173, 182–187

Nexis 115–117

Nineteenth Century Fiction *see*
NCF

non-native speaker 113–114,
123, 127–130, 132–138, 226 *see*
also native speaker

NOW 7–8, 18

O

OB *see* Proceedings of the Old
Bailey

OBC *see* Old Bailey Corpus

OE *see* Old English

OED 42, 82, 88–90, 94, 98, 102,
106, 174–178, 181, 210, 251, 255,
292–293

Old Bailey Corpus 286,
293–294

Old English 169–170, 172–176,
178, 188–189, 193, 199, 251,
280–281, 288

Oxford English Dictionary *see*
OED

P

PDE *see* present-day English

polarity items 178, 188, 191,
193, 228

politeness 57–58, 76, 261

polysemy 6, 8, 68, 79–80, 82,
98, 100, 102, 105, 107, 215

present-day English 34, 161,
169–170, 182, 185, 199, 219,
252, 267, 277–278, 280–281,
283, 297–298

private sector 119–120, 130,
132–134, 136–137 *see also*
public sector

Proceedings of the Old Bailey
262, 267, 286

productivity 81, 172, 199–200,
210, 213

public sector 130, 132, 135–137
see also private sector

R

register 5–6, 61, 74, 171, 226–
227, 245, 285

regression analysis 13, 37,
239–240, 243

response marker 150, 158–160,
162–163

S

SAEMEP 201–205, 210,
213–215, 217

- Selected Authors' Writings in the Early Modern English Period *see* SAEMEP
- Selected Middle English Texts in Print *see* METiP
- semantic prosody 59, 72, 74–75, 169, 171–172, 176, 178, 180, 182, 184–185, 187–193
- semantic similarity 18–19, 231, 235, 242, 244
- semantic vector space 6, 20–21, 23, 230, 235
- semasiological frequency 7–8, 14 *see also* onomasiological frequency
- speaker involvement 144, 152, 156, 158–159
- spelling normalization 34, 42, 51, 56, 61, 63
- spelling variant 32–34, 88, 286–287
- stance 53, 127, 134, 257 *see also* attitude
- stative 154, 162, 284, 291–292 *see also* aspect
- subjectification 25, 172, 175, 190, 193, 268, 272
- subjectivity 143–144, 152, 154, 162, 172, 174–175, 182, 193, 258
- T
- telicity 153, 162, 283–285, 289–291, 292, 298, 302
- Time Magazine Corpus 262
- topic modelling 36, 38–39, 43, 46, 51, 53–54
- V
- VARD 48, 51, 61, 76
- W
- WebCorpLSE 79, 81–82, 109
- WMatrix 39
- WordSmith 118, 173

This book showcases eleven studies dealing with corpora and the changing society. The theme of the volume reflects the fact that changes in society lead to changes in language and vice versa. Focusing on the English language, be it from Old English to the present, or a shorter time span in the immediate past, the contributors in this volume use a variety of corpus methods to address the two patterns of change. The cross-fertilization of cultural studies and corpus linguistics, we hope, is beneficial for both parties, as corpus linguistics offers a vast array of materials and methods to investigate cultural and societal change, while cultural studies provide the theoretical background on which to build our research. The studies included in the present volume illustrate the potential avenues and the merits of combining changing language and changing societies.

ISBN 978 90 272 0543 8



9 789027 205438

JOHN BENJAMINS PUBLISHING COMPANY