Computational approaches to semantic change

Edited by

Nina Tahmasebi

Lars Borin

Adam Jatowt

Yang Xu

Simon Hengchen

Language Variation 6

language
science
press

Language Variation

Editors: Alexandra D’Arcy, John Nerbonne, Martijn Wieling

In this series:

1. Côté, Marie-Hélène, Remco Knooihuizen and John Nerbonne (eds.). The future of dialects.

2. Schäfer, Lea. Sprachliche Imitation: Jiddisch in der deutschsprachigen Literatur (18.–20. Jahrhundert).

3. Juskan, Martin. Sound change, priming, salience: Producing and perceiving variation in Liverpool English.

4. Dellert, Johannes. Information-theoretic causal inference of lexical flow.

5. Zimmer, Christian (ed.). German(ic) in language contact: Grammatical and sociolinguistic dynamics.

6. Tahmasebi, Nina, Lars Borin, Adam Jatowt, Yang Xu and Simon Hengchen (eds.). Computational approaches to semantic change.

ISSN: 2366-7818

Computational approaches to semantic change

Edited by

Nina Tahmasebi

Lars Borin

Adam Jatowt

Yang Xu

Simon Hengchen

language science press

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.).
2021. *Computational approaches to semantic change* (Language Variation 6).
Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/303

ISBN: 978-3-96110-312-6 (Digital)
978-3-98554-008-2 (Hardcover)

ISSN: 2366-7818
DOI: 10.5281/zenodo.5040241
Source code available from www.github.com/langsci/303
Collaborative reading: paperhive.org/documents/remote?type=langsci&id=303

Cover and concept of design: Ulrike Harbort
Typesetting: Simon Hengchen, Lars Borin, Felix Kopecky, Sebastian Nordhoff
Proofreading: Alexandru Craevschi, Amir Ghorbanpour, Annika Schiefner, Eduard S. Lukasiewicz, Gereon A. Kaiping, Jean Nitzke, Jeroen van de Weijer, Kalen Chang, Marten Stelling, Madleine Myers, Peter Petré, Sebastian Nordhoff, Sophie Ellsäßer, Tihomir Rangelov, Viktor Martinovic
Fonts: Libertinus, Arimo, DejaVu Sans Mono, Source Han Serif
Typesetting software: XƎLATEX

Language Science Press
xHain
Grünberger Str. 16
10243 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Contents

Preface

Languages change over time. The process of change is driven, to a large part, by our communicative needs for expressing development in the world around us. While many aspects of language can change, at the semantic level, words can acquire new senses or lose existing ones. They can even, depending on viewpoint, change the senses they represent. We refer to this process as diachronic or historical semantic change. There is rich empirical work on semantic change from historical linguistics, sociolinguistics, and cognitive linguistics. However, computational approaches to historical semantic change have only begun to take shape over the past two decades. It is the latter, computational approaches to semantic change, that are the focus of this edited volume.

The development of the computational field of semantic change has been motivated by a few primary aims. Firstly, the study of semantic change itself, using large-scale digital data, that has been made possible by large-scale digitization efforts. These efforts, hand-in-hand with the rise of digital humanities and social sciences, have resulted in electronic longitudinal text at unprecedented scale. This has provided us with new opportunities for historical investigations of word meaning with the use of computational methods, thus enabling us to test existing hypotheses using data at a much larger scale.

Recently, the inquiry into semantic change has been pursued not only on its own, but also as a basis for other diachronic textual investigation. These include lexicography, culturomics-style studies, temporal classification of unknown texts, and uncovering of document similarities over time.

Next, semantic translations or accessability has been a driving force. With the rise of huge diachronic corpora that are easily accessible to anyone, one motivation has been to make these texts semantically understandable for non-historical linguistic experts. Here, semantic search and temporal information retrieval have been the driving forces.

Finally, semantic change has been used as an application area for modern computational methods. With new, fast, and efficient modeling tools – both topic modeling as well as neural embeddings of different kinds – many researchers have been interested in new problems, and data, to test the limits of computational methods. The time-varying nature of lexical semantics, with many progressing data points, has been one motivation for the rise of interest in computational semantic change.

One of the main challenges for the computational semantic change community so far has been the lack of interaction and collaboration with traditional research and researchers of semantic change in fields like historical linguistics, semantics, typology, and so on. The 1st International Workshop on Computational Approaches to Historical Language Change (LChange'19), held in conjunction with ACL2019, was a first attempt to bring together the international research community around both traditional and computational semantic change, as well as application fields that benefit from semantic change research.[1] The understanding of how our languages behave over time should come from collaboration with, and draw on corresponding efforts within, traditional semantic change research.

Our aim with LChange'19 was to facilitate better collaboration and understanding across fields. This book represents part of that effort, with the main focus on computational semantic change, its applications and open challenges. The scope of this book encompasses a survey of the field of computational semantic change (Chapter 1, Tahmasebi et al. 2021), application fields that benefit, or directly use, semantic change in their research (Chapters 2–4, Vylomova & Haslam 2021, Mahanty et al. 2021, Petersson & Sköldberg 2021), methods for, and investigations into semantic change (Chapters 5–9, Xu & Zhang 2021, Grewal & Xu 2021, Uban et al. 2021, Duan et al. 2021, Perrone et al. 2021). We provide an overview of existing systems and applications where semantic change is incorporated (Chapter 10, Jatowt et al. 2021) and finally, an outlook into the future challenges (Chapter 11, Hengchen et al. 2021).

Even after this book, there are many challenges that remain untackled, and many dimensions along which our field can develop. Bridging the gap between the needs of the widely different applications fields, and the possibilities of (unsupervised) modeling of large scale text, is an important dimension. Solid and shared evaluation frameworks, and evaluation data, is another.

In particular, our field still lacks in-depth analysis of what semantic information each computational model captures, and whether this corresponds to the desired outcome. Because the *optimal* result is highly context dependent, we need to consider the specific needs of the application field in which we are solving problems; for example, the semantic information needed for lexicography will be widely different from what is required in financial, medical, or historical domains. Most evaluation of current computational semantic change show that models capture change of some kind, often in high-dimensional vector spaces, and that this change coincides with certain known properties of our words. However, few

[1] The scope of the workshop was wider and targeted all language change that could be found using textual corpora as a basis.

benefit from *knowing of change* in high-dimensional space without *knowing what* this change corresponds to, be it change in the set of senses associated to the word, or just a lack of interest in the word itself.

We also need to know how much change the different models capture: do they predict change to 90% of the vocabulary and are thus too broad? Do they handle short-term or long-term change? Do they model semantic, syntactic, contextual or cultural change? And do they capture change on different granularity, or only change to a word's main sense?

All of these questions represent opportunities for research, and offer us an exciting future to look ahead to.

References

Duan, Yijun, Adam Jatowt & Masatoshi Yoshikawa. 2021. Structured representation of temporal document collections by diachronic linguistic periodization. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 261–285. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040316.

Grewal, Karan & Yang Xu. 2021. Chaining algorithms and historical adjective extension. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 189–218. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040312.

Hengchen, Simon, Nina Tahmasebi, Dominik Schlechtweg & Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 341–372. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040322.

Jatowt, Adam, Nina Tahmasebi & Lars Borin. 2021. Computational approaches to lexical semantic change: Visualization systems and novel applications. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 311–339. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040320.

Mahanty, Sampriti, Frank Boons, Julia Handl & Riza Batista-Navarro. 2021. Computation of semantic change in scientific concepts: Case study of "circular economy". In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 123–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040306.

Perrone, Valerio, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith & Barbara McGillivray. 2021. Lexical semantic change for Ancient Greek and Latin. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 287–310. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040318.

Petersson, Stellan & Emma Sköldberg. 2021. Semantic change in Swedish – from a lexicographic perspective. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 149–167. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040308.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 1–91. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040302.

Uban, Ana-Sabina, Alina Maria Ciobanu & Liviu P. Dinu. 2021. Cross-lingual laws of semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 219–260. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040314.

Vylomova, Ekaterina & Nick Haslam. 2021. Semantic changes in harm-related concepts in English. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 93–121. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040304.

Xu, Yang & Zheng-sheng Zhang. 2021. Historical changes in semantic weights of sub-word units. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 169–187. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040310.

Acknowledgments

This book has been funded in part by the project *Towards Computational Lexical Semantic Change Detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184), and *Nationella språkbanken* (the Swedish National Language Bank) – a research infrastructure jointly financed by the Swedish Research Council (2018–2024; dnr 2017-00626) and the 10 partner institutions. Yang Xu is funded through an NSERC Discovery Grant RGPIN-2018-05872 and a SSHRC Insight Grant #435190272.

Chapter 1

Survey of computational approaches to lexical semantic change detection

Nina Tahmasebi[a], Lars Borin[a] & Adam Jatowt[b]
[a]University of Gothenburg [b]University of Innsbruck

Our languages are in constant flux driven by external factors such as cultural, societal and technological changes, as well as by only partially understood internal motivations. Words acquire new meanings and lose old senses, new words are coined or borrowed from other languages and obsolete words slide into obscurity. Understanding the characteristics of shifts in the meaning and in the use of words is useful for those who work with the content of historical texts, the interested general public, but also in and of itself.

The findings from automatic lexical semantic change detection and the models of diachronic conceptual change are also currently being incorporated in approaches for measuring document across-time similarity, information retrieval from long-term document archives, the design of OCR algorithms, and so on. In recent years we have seen a surge in interest in the academic community in computational methods and tools supporting inquiry into diachronic conceptual change and lexical replacement. This article provides a comprehensive survey of recent computational techniques to tackle both.

1 Introduction

Vocabulary change has long been a topic of interest to linguists and the general public alike. This is not surprising considering the central role of language in all human spheres of activity, together with the fact that words are its most salient elements. Thus it is natural that we want to know the "stories of the words we use" including when and how words came to possess the senses they currently have as well as what currently unused senses they had in the past. And while

Nina Tahmasebi, Lars Borin & Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 1–91. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040302

some examples are commonly known, like *gay* having meant 'happy' in the past, the fact that *girl* used to mean 'young person of either gender' is unknown to many. Professionals and the general public are interested in the origins and the history of our language as testified to by numerous books on semantic change aimed at a wide readership.

Traditionally, vocabulary change has been studied by linguists and other scholars in the humanities and social sciences with manual, "close-reading" approaches. While this is still largely the case inside linguistics, recently we have seen proposals, originating primarily from computational linguistics and computer science, for how semi-automatic and automatic methods could be used to scale up and enhance this research.

Indeed, over the last two decades we have observed a surge of research papers dealing with detection of lexical semantic changes and formulation of generalizations about them, based on datasets spanning decades or centuries. With the digitization of historical documents going on apace in many different contexts, accounting for vocabulary change has also become a concern in the design of information systems for this rapidly growing body of texts. At the same time, as a result, large scale corpora are available that allow the testing of computational approaches for related tasks and that provide quantitative support to proposals of various hypotheses.

Despite the recent increase in research using computational approaches to investigate lexical semantic changes, the community is in critical need of an extensive overview of this growing field. The aim of the present survey is to fill this gap. While we were preparing this survey article, two related surveys appeared, illustrating the timeliness of the topic.[1] The survey by Kutuzov et al. (2018) has a narrower scope, focusing entirely on diachronic word embeddings. The broader survey presented by Tang (2018) covers much of the same field as ours in terms of computational linguistics work, but provides considerably less discussion of the connections and relevance of this work to linguistic research. A clear aim in preparing our presentation has been to anchor it firmly in mainstream historical linguistics and lexical typology, the two linguistic subdisciplines most relevant to our survey. Further, the application of computational methods to the study of language change has gained popularity in recent years. Relevant work can be found not only in traditional linguistics venues, but can be found in journals and conference proceedings representing a surprising variety of disciplines, even outside the humanities and social sciences. Consequently, another aim of this survey has been to provide pointers into this body of research, which often

[1] This survey is an updated and published version of the survey presented by Tahmasebi et al. (2018).

utilizes datasets and applies methods originating in computational linguistics research. Finally, our main concern here is with computational linguistic studies of vocabulary change utilizing empirical diachronic (corpus) data. We have not attempted to survey a notable and relevant complementary strand of computational work aiming to simulate historical processes in language, including lexical change (see Baker 2008 for an overview). We also leave out of consideration work utilizing digitized historical dictionaries as the primary data source (e.g., Xu et al. 2017, Ramiro et al. 2018, Cathcart 2020). While historical text digitization initiatives are often undertaken by public cultural heritage institutions such as national libraries, historical dictionaries are as often as not commercial ventures which makes them both very scarce and often not freely accessible in a way which would allow reproducibility of experiments, let alone release of enriched versions of the dictionaries.[2]

The work surveyed here falls into two broad categories. One is the modeling and study of DIACHRONIC CONCEPTUAL CHANGE (i.e., how the meanings of words change in a language over shorter or longer time spans). This strand of computational linguistic research is closely connected to corresponding efforts in linguistics, often referring to them and suggesting new insights based on large-scale computational studies, (e.g., in the form of "laws of semantic change"). This work is surveyed in two sections, one section on word-level change in Section 3, and one on sense-differentiated change in Section 4. The word-level change detection considers both count-based context methods as well as those based on neural embeddings, while sense-differentiated change detection covers models based on topic modeling, clustering, word sense induction, and – the most recent development – contextualized embeddings.

The other strand of work focuses on LEXICAL REPLACEMENT, where different words express the same meaning over time. This is not traditionally a specific field in linguistics, but it presents obvious complications for access to historical text archives, where relevant information may be retrievable only through an obsolete label for an entity or phenomenon. Because successful approaches to semantic change over longer time scales are strongly dependent on the possibility to first resolve lexical replacements, we cover this body of work in Section 5.

The terminology and conceptual apparatus used in works on lexical semantic change are multifarious and not consistent over different fields or often even within the same discipline. For this reason, we provide a brief background synopsis of relevant linguistic work in Section 2.

[2] For instance, according to the website https://ht.ac.uk/terms/, accessed April 4th, 2021, the *Historical Thesaurus of English* that was used for the studies of Xu et al. (2017) and Ramiro et al. (2018) is available for research by agreement only and only on quite specific conditions, to a limited number of research projects at the same time.

Much current work in computational linguistics depends crucially on (formal, automatic, quantitative, and reproducible) evaluation. Given the different aims of the surveyed research, evaluation procedures will look correspondingly different. We devote Section 6 to a discussion of general methodological issues and evaluation.

We end with a summary of the main points garnered from our literature survey, and provide a conclusion and some recommendations for future work (Section 7).

We believe our survey can be helpful for both researchers already working on related topics as well as for those new to this field, for example, for PhD candidates who wish to quickly grasp the recent advances in the field and pinpoint promising research opportunities and directions.

2 Linguistic and computational approaches to vocabulary change

2.1 Terminological and conceptual prelude

The study of how meaning – including lexical meaning – is expressed and manipulated in language is pursued in a number of scientific disciplines, including psychology, (cultural) anthropology, history, literature, philosophy, cognitive science, and in linguistics and computational linguistics. These all construe the problems involved in studying linguistic meaning in different ways, for different purposes, and consequently conceptualize this field of inquiry differently, with concomitant differences in terminology. Drawing on partly common origins, they unfortunately often use the same terms, yet with different meanings.

Our primary frame of reference in this chapter is provided by relevant work in (general) linguistics, being the field offering the theoretically and empirically best-grounded view on the phenomena under discussion here. In particular, in studying meaning in language, linguistics takes a broad cross-linguistic perspective, which is typically lacking in the other disciplines addressing this question.

Because many of the terms found in discussions of lexical change are not used in the same way by all authors, we start out by defining our use of some central terms. In order to discuss linguistic semantics and semantic change over time, we need to distinguish the following notions. linguistic form or linguistic substance is the physical manifestation of language: linguistic expressions formed using sound, writing, or sign(ed language). In addition, linguistic form is normally taken to include certain structural aspects of language expressions, such as parts of speech, inflectional paradigms, dependency trees, and so on. MEANING OR SENSE is information – in a wide sense – conventionally connected with (or conveyed by) the forms. It is essentially thought of as something residing in the minds of the language users. The sense is what a dictionary definition aims to capture. Linguistic meaning is generally considered to exhibit at least two aspects. DENOTATION OR DENOTATIVE MEANING corresponds to the "neutral" information content. CONNOTATION OR CONNOTATIVE MEANING refers to attitudinal or sentiment-conveying aspects. The English words *thrifty* and *stingy* have by and large the same denotation but different connotations.

Finally, linguistic meaning connects language to the extralinguistic realm: to the actual world and also to imagined situations. Here, the terminology becomes more motley, and for our purposes in this chapter it will suffice to note that the relation of linguistic meaning to extralinguistic reality can be seen as indirect – mediated by mental CONCEPTS[3] – or direct – the case of proper nouns, which refer directly. The main function of a personal name like *Faith* is to pick out an individual and the fact that the word also corresponds to a common noun is of no import in this case,[4] and does not help us identify the individual in question.

Students of human linguistic behavior and language have been investigating and discussing the nature of these notions and their relationships for millennia, so this brief introduction cannot do justice to all the complexities involved. Rather, we have tried to summarize briefly what we understand as a view broadly shared among linguists, and only to the extent necessary for the present survey.

In this chapter, the linguistic forms in focus are LEXICAL ITEMS, i.e., words (or multiword expressions) that are not *semantically* decomposable into smaller parts.[5] Among the lexical items we also include proper nouns and function words. Interchangeably with lexical item we will also say "word", intending this term also to apply to multiword expressions.

[3] Some authors make no distinction between "meaning" and "concept", and both terms unfortunately have many – sometimes mutually incompatible – uses in the literature. "Concept" is especially treacherous, since it is treated – sometimes explicitly defined – as a term, but with widely differing content in different contexts. E.g., the "concepts" of conceptual historians seem to actually be simply words, reflected in statements about the "changing meaning of concepts" (Richter 1996), which makes their undertaking tantamount to (a kind of) etymological study. We will use the two terms – sparingly – interchangeably here, with the understanding that neither term is well-defined.

[4] Thus, the personal name *Faith* will not be "translated" into Russian *Vera* or Finnish *Usko*, both of which in addition to being personal names are also common nouns literally meaning 'faith, belief' (and the Finnish correspondent is actually a male name).

[5] Although they will often be *formally* decomposable, the semantics of the whole is not computable from that of the parts.

Note that lexical items are not the same thing as TEXT WORDS. A lexical item in our usage of this term corresponds roughly to what is often called LEXEME in lexicography (e.g., Matthews 1974), basically what we understand as an entry – a word or multiword expression – in a conventional dictionary, referring through a CITATION FORM or LEMMA to a bundle of formal characteristics, including at least a part of speech and possibly a set of inflected forms, which make up the text words subsumed by the lexical item. The inflectional pattern, while an important clue to lexemehood in many languages, is not so salient in English, where generally lemma and part of speech are sufficient to uniquely identify a lexical entry, but an example could be *stick* (v). It corresponds to two such lexical units: one with the past form *stuck* 'to pierce, to fasten, etc.' and another with the past form *sticked* 'to furnish (a plant, vine, etc.) with a stick or sticks in order to prop or support'. Another example: *die* (n), with the plural form *dies* 'a cutting or impressing tool' or *dice* 'small cube with numbered sides used in games'.

We will refer to the combination of a lexical item and a particular recognized meaning of that lexical item as a WORD SENSE. Thus, both *bank* (n) '(a kind of) financial institution' and *bank* (n) 'extended shallow portion of sea or river floor' are word senses according to this definition, as are *moose* (n) 'a kind of (game) animal' and *moose* (n) 'meat of this animal used as food'.

The relationship between forms and meanings is many-to-many, so one form may be used to express more than one meaning, and, conversely, the same meaning can be expressed by more than one form. The former configuration will be consistently referred to as POLYSEMY (or COLEXIFICATION[6]) even when some lexicographical traditions would distinguish it from HOMONYMY. This distinction is hard or impossible to make categorically (Apresjan 1974, Murphy 2003, Riemer 2010, Wishart 2018), so we have not attempted to make it.[7] The latter configuration is known as (NEAR) SYNONYMY, and, depending on its definition in a particu-

[6] This is a more neutral term often encountered in the lexical typological literature intended to cover both polysemy and homonymy (e.g., François 2008, Östling 2016).

[7] According to Apresjan (1974) we should recognize polysemy (as opposed to homonymy) when two senses of a word exhibit non-trivial common components in their definitions. However, he does not discuss how to ensure intersubjective agreement on definitions, which makes this criterion less than exact. Similarly for the "technical definition of concept" (where "concepts" correspond to homonymous – main – senses of a lexeme) provided by Cooper (2005: 235; emphasis in the original): "Two meanings of a given word correspond to the same *concept* if and only if they could inspire the same new meanings by association." Again, there is no indication in the article of how this definition could be operationalized to ensure intersubjective agreement. This is not to deny that lexeme meanings can be seen as hierarchically organized or that the intuitions behind the cited statements are well-founded, but simply to recognize that there are no straightforwardly applicable mechanical criteria for distinguishing polysemy from homonymy, and also – which Apresjan acknowledges – that in reality this is not a dichotomy,

lar lexicographical tradition, it may be seen as frequent (as in a wordnet) or next to non-existent (Cruse 1986, Ci 2008/1987, Murphy 2003, Riemer 2010).

While the form units – the words – are comparatively easy to identify in language, word senses are notoriously difficult to isolate. Much of the work surveyed in this chapter takes a published lexicon as providing the canonical sense set, the gold standard by which to judge system accuracy. While this is a practical solution for many purposes, it also, in effect, ignores a host of difficult theoretical and methodological questions. For the purposes of this survey, we do not take a stand on precisely how word senses are defined and identified, but we do note that some of the approaches represented in the surveyed work have the potential to throw light on these questions; see below.

2.2 Linguistic studies of lexical change

To a linguist, the topic of this chapter would fall under the rubric of HISTORICAL-COMPARATIVE LINGUISTICS or DIACHRONIC LINGUISTICS. This is a branch of general linguistics that concerns itself with how languages change over time and with uncovering evidence for genetic relations among languages (Anttila 1972, Campbell 2004, Joseph & Janda 2003). This linguistic subfield has a long history, antedating by a century or so the birth of modern SYNCHRONIC LINGUISTICS. The latter by and large emerged in the early twentieth century in no small measure as a reaction against the predominant historical orientation of mainstream linguistics of the time.

Even if now relegated to a more modest position within the language sciences, historical-comparative linguistics is very much alive and an active branch of linguistic research. For this reason it is interesting to elucidate how it interacts, or could interact, with the computational linguistics research surveyed here.

2.2.1 Lexical change, semantic change, grammaticalization, and lexical replacement

The phenomena addressed in the works surveyed in this chapter (i.e., historical developments in the vocabulary of a language or languages) are studied by historical linguists under the headings of LEXICAL CHANGE, SEMANTIC CHANGE, GRAMMATICALIZATION, and LEXICAL REPLACEMENT.

but rather a cline. Consequently, some of the methods discussed in this survey article could in fact be applied also to the problem of teasing out hierarchical relationships among word senses of the same lexeme.

In linguistic literature, the term lexical change unfortunately is used in two senses. In the sense used here, it is a general cover term for all kinds of diachronic changes in the vocabulary of a language or languages. The other common usage is a hyponym of this, referring to new forms entering or leaving the language, i.e., loanwords and neologisms of various kinds, and obsolescing words, respectively.

Lexical replacement refers to a lexeme being ousted by another synonymous lexeme over time, as when *adrenaline* is replaced by *epinephrine*. A particular form of lexical replacement which has received a fair amount of attention in computational linguistics but which is generally not studied at all by historical linguists is NAMED ENTITY CHANGE.[8]

Semantic change or semantic shift is the normal term for the special case of lexical change where an existing form (a lexeme) acquires or loses a particular meaning, i.e., increasing or decreasing polysemy (Traugott & Dasher 2001, Fortson 2003, Newman 2016, Traugott 2017). An example are the oft-cited changes whereby on the one hand an earlier English word for a particular kind of dog became the general word for ‘dog’, and, on the other, the earlier general word for ‘dog’ – whose modern reflex is *hound* (n) – is now used for a special kind of dog.

There are two complementary approaches adopted by linguists to the study of the lexicon. Lexical items can be studied from the ONOMASIOLOGICAL point of view, investigating how particular meanings (or concepts) are expressed in a language. The Princeton WordNet (Fellbaum 1998) is an onomasiologically organized lexical resource, as is, e.g., *Roget’s Thesaurus* (Roget 1852). The more common SEMASIOLOGICAL approach takes linguistic forms – words and multiword expressions – as its point of departure and investigates which meanings they express. Conventional dictionaries are semasiologically organized.

[8] This is most likely because, strictly speaking, named entity change does not involve word senses at all (see above). However, the *etymology* of names – in particular place names – plays an important role in historical linguistics, where it is studied under the label of TOPONYMY, as a clue to determining prehistorical linguistic geography and population movements. For example, the fact that the city names *Dresden* and *Leipzig* both have a recognizable Slavic origin is taken to confirm a more westerly extension of Slavic speakers in earlier times in present-day Germany. This is also indicated by historical records. It is also true that names can be the basis for general vocabulary, in other words, the etymology of a non-name must sometimes make reference to a name. For example, *bedlam*, from the (nick)name of a psychiatric hospital in London, or the (Chilean) Spanish verb *davilar* ‘to botch things up royally’, from the surname of Juan Pablo Dàvila, an infamous spectacularly inept financial trader (https://www.improbable.com/ig/winners/#ig1994). Finally, a cultural taboo against naming the dead may lead to avoidance of words sounding like the name of a recently deceased person, replacing them with, e.g., loanwords (Alpher & Nash 1999: 8f).

Studies of semantic change adopt the semasiological perspective, whereas works on other forms of lexical change generally have an onomasiological focus.

Grammaticalization (Hopper & Traugott 1993, Heine & Kuteva 2002, Smith 2011) denotes a particular kind of semantic change, where content words turn into function words and ultimately into bound grammatical morphemes. One example is the French preposition *chez* 'at, with', developed from the Latin noun *casa* '(small) house, cottage'.[9]

In both semantic change and grammaticalization, the form is thus fixed – modulo historical sound shifts[10] – while its content changes.

The term ETYMOLOGY refers to the scientific investigation of the origin and history of lexical items, whose development may include both onomasiological and semasiological aspects (Malkiel 1993, Anttila 1972, Mailhammer 2015). In fact, these aspects interact in a natural way, and are perhaps best thought of as different views on a unitary phenomenon, viz. lexical change.

2.2.2 Theoretical and methodological aspects of the linguistic study of lexical change

A central activity in the linguistic study of vocabulary change is the description of individual changes in the vocabulary of a language or group of related languages. The concrete outcome of this research is the etymological article or dictionary.

As its name indicates, general linguistics studies language as a universal phenomenon, and collecting data about individual languages is thought of as contributing to this goal. Consequently, an important concern of this field of inquiry is the generalization of sets of observed individual lexical changes into types and classes of changes, valid for human languages in general. This includes uncovering universal or general directional tendencies – "laws" – of semantic change, such as person-part > enclosing person-part (e.g., 'mouth' > 'face'), but not the opposite (Wilkins 1996), many individual grammaticalization paths and, more generally, the assumed unidirectionality of grammaticalization (Heine & Kuteva 2002, Smith 2011).

The common event of adding a word sense to the vocabulary of a language can be accomplished in several different ways. These are, by borrowing, coining

[9] See http://www.cnrtl.fr/definition/chez.

[10] That is, Latin *casa* and French *chez* count as the same word, even though they do not in fact share a single speech sound (*casa* sounded more or less as expected – [ˈkasa] – while *chez* is pronounced [ʃe]), since the latter is derived from the former by regular historical sound changes.

a new word *ex nihilo* (rare) or using the word-formation machinery of the language, or finally – and commonly – adding a word sense to an existing lexeme. The latter can again be achieved by, for example, GENERALIZATION OR BROADENING (English *dog* 'a kind of dog' > 'dog')[11] and SPECIALIZATION OR NARROWING (English *hound* 'dog' > 'a kind of dog'). Other types of semantic change have their origin in METAPHOR, as in the *foot* of a mountain or the *head* of a state; in METONYMY, for example, the development where *bead*, a word originally meaning 'prayer', acquired its current meaning from the use of a rosary while praying; and in ELLIPSIS, as *mobile* and *cell* from *mobile phone* and *cell phone*, respectively. For a more detailed oveview of (lexical) semantic change and how this phenomenon has been studied by linguists, see Urban (2015). Finally, a lexeme in one language may add a sense by mirroring a polysemy in another language, a form of loan translation. For example, the Swedish verb *suga* 'to suck' has acquired a recent new sense 'to be unpleasant, inferior, etc.' borrowed from English. From this it follows that semantic change typically involves polysemy or colexification. Crucially, even cases of seemingly complete sense change in a lexeme are thought to involve an intermediate (unattested) polysemous stage: A > A+B > B, or A > A+b > a+B > B, where A/a and B/b are senses related by some regular mechanism of sense change and caps indicate a dominant sense. Thus, variation in the language community in the distribution of these colexified senses is what ultimately drives semantic change (Bowern 2019).

The activities of broadly characterizing and classifying vocabulary changes overlap significantly with another linguistic subdiscipline, namely LEXICAL TYPOLOGY (Koptjevskaja-Tamm 2008, 2012, Koptjevskaja-Tamm et al. 2016). This is also referred to as SEMANTIC TYPOLOGY (Riemer 2010), whose aims are to elucidate questions such as "how languages categorize particular domains (human bodies, kinship relations, colour, motion, perception, etc.) by means of lexical items, what parameters underlie categorization, whether languages are completely free to "carve up" the domains at an infinite and arbitrary number of places or whether there are limits on this, and whether any categories are universal (e.g., 'relative', 'body', or 'red')" (Koptjevskaja-Tamm et al. 2016: 434). These questions are relevant to classificatory activities, since universal restrictions on or tendencies of lexicalization will determine which semantic changes are possible or likely, as opposed to impossible or unlikely.

However, as Anttila (1972: 148) observes, "labeling before-after relations […] does not explain anything; it just states a fact", and a central goal of linguistics is to explain linguistic phenomena. Hence, a third kind of activity is the search for

[11] Generalization is also considered to make up an important initial stage of grammaticalization (Smith 2011).

enabling factors and, ultimately explanations for the observed changes and regularities of change formulated on the basis of broad cross-linguistic comparison.

In their search for explanations of lexical change, linguists have proposed some factors that seem to play a role in lexical change, as (proximal or distal) causes or as enabling or constraining mechanisms. Material and immaterial culture are almost always mentioned in this connection. In order to be able to talk about new objects, phenomena, and practices, we need new vocabulary, so the argument goes. At one point, historical linguists saw this as a – or even the – major driving force behind lexical change, a point of view forcefully argued by the *Wörter und Sachen* 'words and things' school active at the beginning of the 20th century (Meringer 1912).

Other potentially influencing factors, which have been discussed in the linguistic literature, are human physiological and cognitive characteristics (e.g., in relation to color vocabulary), systematic sound symbolism/onomatopoeia (Erben Johansson et al. 2020), the size of the language community, language contact, and the presence of large numbers of L2 speakers, among others. For example, Ellison & Miceli (2017) adduce linguistic and psycholinguistic evidence that bilinguals speaking closely related languages develop a cognitive bias against recognizably shared word forms (termed "doppels" by Ellison & Miceli 2017), which they argue accelerates lexical change.

2.3 Historical-comparative linguistics meets computational linguistics?

When historical linguists started to use computers more than half a century ago, their primary focus was initially on modeling sound change as formal rule systems, in order to check that postulated changes yield the expected outcome, or to reverse the changes to produce putative proto-forms from modern forms (e.g., Hewson 1973, 1974, Johnson 1985, Borin 1988, Lowe & Mazaudon 1994). In more recent times and coinciding with the statistical and machine-learning emphasis characterizing present-day computational linguistics, massively multilingual datasets have been employed for genealogical classification of languages (Brown et al. 2008).

In the linguistic subfield of corpus linguistics,[12] the increasing availability of large historical text sets has spurred corpus-based work on historical semantics and pragmatics (Ihalainen 2006, Taavitsainen & Fitzmaurice 2007, Allan & Robinson 2011). This work is typically semasiological and particularistic in spirit, tak-

[12] Corpus linguistics is related to computational linguistics but often surprisingly separate from it. The two fields do share an interest in applying computational methods to language, but at the same time they differ crucially in their primary aims.

ing as its point of departure particular words – given a priori – and endeavoring to track their shifting semantics over time (e.g., Sagi et al. 2011, Kerremans et al. 2011). The only efforts we are aware of in this area to address the problem in a more general way do so only indirectly. Koplenig (2017a) and Degaetano-Ortlieb & Strötgen (2017), for example, describe computational methods for identifying changing word usages over time in diachronic text, but it is reasonable to assume, ceteris paribus, that these changes often (or always) will reflect changing semantics of the forms thus identified.

While some of the work described and discussed in the present survey has not been directly motivated by linguistic research questions, the authors of these works often indicate the potential usefulness of their results to linguistics. We believe that computational approaches to lexical and semantic change have the potential to provide a genuinely novel direction for historical linguistics. However, this is not likely to happen without these authors paying more attention to the theoretical and methodological assumptions of current historical linguistics, an awareness sometimes lacking in the work surveyed. For linguists to take notice of this work, it needs to show awareness of the state of the art of diachronic linguistics and argue in terms understandable to a linguistic audience.

In this connection, a central methodological question will be REPRESENTATIVENESS. During the rapid growth phase of corpus linguistics in the 1970s and 1980s, representativeness was a much discussed concern (e.g., Atkins et al. 1992, Biber 1993, Clear 1992, Johansson 1994), the issue of course being the question if we will be able to say anything meaningful about our actual object of study, the language, when investigating the corpus. The question remains, but tends to be rarely addressed in the computational linguistics literature, one notable exception being the work of Koplenig (2016, 2017a).

In diachronic studies, the demands for representativeness are exacerbated by the requirement to compare two or more temporal language stages. We must ensure that all investigated time-slice subcorpora are equally representative of their respective language stages. Linguistic differences between the subcorpora must not be caused by some confounding extralinguistic factor. An example may make this more concrete. Underwood (2019: Ch. 4) – a literary scholar – presents a study of "gendered language": words used to portray feminine and masculine characters in English-language fiction in the period 1840–2000. First, and importantly to our example, the study shows that there are clear demonstrable differences in terms of the words used by authors for depicting masculine and feminine characters and their actions, although the differences grow smaller over the course of the twentieth century. However, the study also reveals some relevant additional facts, namely

- "a fairly stunning decline in the proportion of fiction writers who were women from the middle of the nineteenth century to the middle of the twentieth […] from representing almost half the authors of fiction to less than a quarter" (Underwood 2019: 133); and

- that over the same period, "[w]omen are constantly underrepresented in books by men" (Underwood 2019: 127).

These two facts together could lead to words used specifically to describe feminine characters exhibiting a significant shift in distribution over time in such a diachronic fiction material, which could be interpreted as semantic change.

On the other hand, the most crucial awareness is simply this: "Knowing that your corpus is unbalanced is what counts. It would be shortsighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as 'unreliable' or 'irrelevant' simply because the corpus used cannot be proved to be 'balanced'." (Atkins et al. 1992: 6). In particular with historical data, it may not even be possible to achieve balance in the sense expected from a modern corpus.

As discussed above, lexical change can be seen as a special case of lexical variation, which in turn can be attributable to many different linguistic and extralinguistic factors. In other words, we see the task of establishing that we are dealing with variants of the same item (in some relevant sense) – items of form or content – as logically separate from – and logically prior to – establishing that the variation is classifiable as lexical change.

Investigation of lexical change is further complicated by the fact that – as just noted – observed variation in lexical form between different text materials need not be due to diachronic causes at all, even if the materials happen to be from different time periods. Linguists are well aware that even seen as a synchronic entity, language is full of variation at all linguistic levels. In spoken language, this kind of variation is the norm. Words have a wide range of pronunciations depending on such factors as speech rate, register/degree of formality, phonetic and phonological context, etc. If the language has a written form, some of this variation may be reflected in the orthography, but orthography may also reflect ambiguous principles for rendering some sounds in writing, as when /s/ can be written alternatively (at least) with ⟨s⟩, ⟨c⟩, ⟨z⟩ and ⟨ps⟩ in Swedish. Spelling principles – if standardized at all, which often is not the case in older texts – may change over time independently of any changes in pronunciation ("spelling reforms"), and in such situations written texts may exhibit a mix of the older and newer orthography. Finally, in many modern text types we find a large number of spellings which deviate from the standard orthography (Eisenstein 2015).

A fundamental question underlying all work on semantic change is the problem of identifying like with like, or – on the form side – classifying text words under relevant lexical units, and – on the content side – identifying and grouping relevant senses.

Although often trivial, even the former task is complicated by the existence of multiword expressions, the need for word segmentation (in speech and some writing systems), and – a fortiori in a diachronic context – language variation, which may be purely orthographic, both synchronically and diachronically, as well as a reflection of sound change in the diachronic setting.[13]

The latter task is widely recognized to be unsolved, and possibly not even amenable to finding one solution in that there will not be one canonical sense set for a particular language, but several sets depending both on their intended use (Kilgarriff 1997), on particular analytical traditions ("lumpers" vs. "splitters"), and even on individual idiosyncrasies.[14] In this context work such as that surveyed here can make a real contribution, by putting the identification of senses on a much more objective footing, and also allow for different sense granularities for different purposes by adjusting model parameters (Erk 2010).

On a more basic level, these questions are intimately related to some of the basic theoretical and methodological conundrums of linguistics, such as the nature of words (Aikhenvald & Dixon 2002, Haspelmath 2011), of concepts (Murphy 2002, Wilks 2009, Riemer 2010) and their relation to word senses (Cruse 1986, Kilgarriff 1997, 2004, Hanks 2013).

Generally speaking, training in (historical) linguistics prepares researchers to take such confounds and caveats into account, giving them a fair idea of which the crucial non-relevant variables are likely to be, and, importantly, how to design investigative procedures which "short-circuit" such variables. Lack of such training of course comes with the risk that experiments will be poorly designed or their results misinterpreted.

In the final count, however, the computational methods surveyed in this chapter represent a genuinely novel approach to addressing many research questions of historical linguistics, and linguists must be prepared to assimilate the methods

[13] Orthography interacts in intricate ways with language change. Since spelling is often conservative, it may provide hints about earlier, pre-sound change forms of words, such as written-word initial ⟨kn-⟩ in English (e.g., *knight*), which may help us to see connections among lexical items which have otherwise been obscured by sound change. A (sporadic) case such as English ⟨discreet⟩ vs. ⟨discrete⟩ – where two spelling variants of the same original item (still pronounced identically) parted ways in the late 16th century (https://www.dictionary.com, s.v. *discreet*) – will serve as concrete evidence of polysemy, although not of course in an exclusively written-language setting.

[14] Or on completely extraneous factors, such as budget constraints (Lange 2002).

at least to some extent in order to grasp the implications of the results. Thus, if these methods are to make an impact on research in historical linguistics – as we think they could – a conceptual shift is most likely required in both camps.

2.4 Computational studies of lexical change: A classification

Relating the main kinds of lexical change which have been considered in computational linguistics to those discussed in historical linguistics, we note that there is no neat one-to-one correspondence. The study of *semantic change* looms large in both fields and by and large focuses on the same kinds of phenomena, but in computational work, this is typically combined with a study of gain and loss of lexemes (i.e., *lexical change* in the narrower sense), since these phenomena are uncovered using the same computational methods. This could be said to constitute a consistent focus on the conceptual side of the vocabulary, which however is not normally present in historical linguistics and consequently not given a label. In this survey, we refer to it as DIACHRONIC CONCEPTUAL CHANGE, i.e. change in the set of lexical meanings of a language. We propose this term as a superordinate concept to semantic change. Diachronic conceptual change takes the view of all senses and word-sense allocations in the language as a whole. This includes a new word with a new sense (e.g., neologisms like *internet* with a previously unknown sense) as well as an existing word with a new sense (*gay* firstly receiving a 'homosexual' sense, and later more or less losing its 'cheerful' sense), because both of these add to the set of senses available in the language. Diachronic conceptual change also allows for changes to the senses themselves, the line between actual meaning change and usage change is blurry here. Examples include the *telephone* that is a 'device for conveying speech over a distance', but that is now also used for spread of communication, and increasingly as a 'personal device used for photography, scheduling, texting, working', and so on.

Further, the specific phenomena of *lexical replacement* (including *named entity change*) and its generalized version TEMPORAL ANALOGY have been subject to many computational linguistic studies. Examples include the placename *Volgograd* that replaced *Stalingrad*, which in its turn earlier had replaced *Tsaritsyn* (named entity change), *foolish* that replaced *nice* for the 'foolish' sense of the latter word (lexical replacement), and *iPod* that can be seen as a temporal analog of a *Walkman*. The change classes and their ordering as they are being studied from a computational perspective are shown in Table 1.1, and different types of semantic change are shown in Table 1.2.

Table 1.1: Change types and their organization considered from a computational perspective

Lexical semantic change	
Lexical change	Diachronic conceptual change
Lexical replacement	Semantic change (new allocation of existing words and senses)
Named Entity change	Novel form to denote a known entity
Role changes	New words with completely new word sense
Temporal analogy	New word with a new but existing sense
	Changes to existing senses

Table 1.2: Change types investigated in the surveyed literature (ws = word sense)

Change type	Description
Novel word	a new word with a new sense
Novel word sense	a novel word sense that is attached to an existing word
Novel related ws	a novel word sense that is related to an existing sense
Novel unrelated ws	a novel word sense that is unrelated to any existing sense
Broadening	a word sense that is broader in meaning at a later time
Join	two word senses that exist individually and then join at a later time
Narrowing	a word sense that is broader in meaning at an earlier time
Split	a word sense that splits into two individual senses at a later time
Death	a word sense that is no longer used
Change	any significant change in sense that subsumes all previous categories

3 Computational modeling of diachronic semantics

In 2008, the first computational models in the field of diachronic semantics appeared. First a model paper differentiating between different kinds of lexical semantic change (Tahmasebi et al. 2008), while the first empirical study was presented a year later by Sagi et al. (2009). After that, a few papers per year were presented until the first use of neural embeddings as a basis for modeling meaning (Kim et al. 2014). Since then, the field has seen an increasing number of papers per year. In 2019, a first tutorial was given on the topic (Eisenstein 2019), and the first international workshop on computational approaches to historical language change (LChange’19) was held during ACL2019, (Tahmasebi et al. 2019) where another 14 papers were devoted to the topic out of a total of 34 papers devoted to all aspects of language change.[15] In 2020, the first SemEval task on *unsupervised lexical semantic change detection* was held on four languages (Schlechtweg et al. 2020) and soon followed by the EVALITA 2020 *diachronic lexical semantics* (DIACR-Ita) task on Italian (Basile et al. 2020).

In our survey work, we will split the modeling of diachronic semantics into two sections: In this section we cover word level change detection and in the next section sense-differentiated methods. Methods surveyed in both sections rely on semantic modeling of words and the foundation for all methods lie in the well-known distributional hypothesis: "You shall know a word by the company it keeps" (Firth 1957: 11) (pure frequency methods excluded). Regardless of whether pure co-occurrence computing, or contextualized embedding methods are used, a word's meaning or senses rely on the context in which they appear in a written corpus.

Table 1.3: Structure of the two sections on diachronic conceptual change

Word-level sense change (§3)	Sense-differentiated sense change (§4)
§3.1 Co-occurrence-based methods	§4.1 Topic-based models
§3.2 Static Neural Embeddings	§4.2 WSI-based models
§3.3 Dynamic word embeddings	§4.3 Deep contextualized embeddings
§3.4 Laws of sense change	§4.4 Aligned corpora
§3.5 Related technologies	§4.5 Comparison

The methods presented in this section aim to capture diachronic conceptual change from a computational perspective and rely on different embedding techniques for representing words. While the papers surveyed in Section 3.2 feature (static or type-based) neural embeddings, the papers surveyed in Section 3.1.1 employ co-occurrence vectors in different ways.[16] All methods in this section represent all senses of a word using a single representation, that is, no sense discrimination or induction takes place. Within the subsections, we have ordered the papers in diachronic order. The majority of the papers evaluate some aspects in a systematic manner, while many results are presented in an anecdotal fashion,

[15] https://languagechange.org/events/2019-acl-lcworkshop/

[16] Contextualized methods, like ELMo and BERT, produce token embeddings specific to the context in which a word appears. These have the discriminatory power to separate into senses and are surveyed in Section 4, though there are examples that average across all usages and thus fall under word-level sense change (Martinc, Kralj Novak, et al. 2020).

often not accompanied by explicit judgments by the author(s). For a systematic evaluation and comparison of some of the methods presented below, we refer to Schlechtweg et al. (2019).[17]

3.1 Co-occurrence-based methods

Most of the methods presented in this section make use of co-occurrence information, and first build co-occurrence matrices. In a co-occurrence matrix, the information in a corpus is summarized to capture which words occur in close proximity in the text. Each row corresponds to a word, e.g., *happy*, and the columns correspond to the words in the vocabulary. So if there is a vector of *happy* as follows *happy* = (0, 1, 4, …) that means that *happy* does not co-occur with the 1st word in our vocabulary, it occurs once with the 2nd word, four times with the 3rd word, and so on. Each vector (i.e., row in the matrix) has |V| number of elements. These matrices tend to be large (|V|*|V| size, where |V| is the size of the vocabulary) and only few of the elements are nonzero, that means, most words co-occur with few other words. Therefore, many tricks are used to reduce the size of the co-occurrence matrix, and to increase the information. Firstly, few use all the words that appear in a corpus: for example, many use the top (i.e., most frequently occurring) 10,000 text words (or lemmas). Secondly, the majority use pointwise mutual information (PMI) scores of different kinds (local, global or positive), rather than raw frequency scores for co-occurrence strength (Bullinaria & Levy 2012, Levy et al. 2015, Turney & Pantel 2010). These are measures of association given evidence in the underlying corpus. Finally, the number of elements in each vector can be radically reduced using singular value decomposition (SVD) (Eckart & Young 1936), which reduces the length of each vector to a fixed dimension, for example 300, while keeping the most important information from the original matrix. After SVD, however, the values in each column lose their interpretability; they no longer state how often word w co-occurs with word i, for each position $i = 1, \ldots, |V|$. This abstraction and in essence, summarization of information, has often turned out to significantly outperform raw co-occurrence matrices.

Similarity is measured almost exclusively using cosine similarity. Rodda et al. (2017) make use of second order similarity rather than work on first order similarity. Kahmann et al. (2017) use a rank series and compare differences in rank over

[17] In early 2020, the first SemEval task on unsupervised lexical semantic change detection was launched in which manually annotated, sense-differentiated gold labels were released for four different languages. While many systems participated in the task, none of the papers in this survey have used these testsets for evaluation. For a summary of the task and the participating systems, we refer to Schlechtweg et al. (2020).

time. The most distinctive are the works by Basile et al. (2016) who use random vectors to represent each word together with context information, and Tang et al. (2013) who use contextual entropy and reduce dimensions on the fly rather than applying SVD as post-processing.

3.1.1 Context vectors

Sagi et al. (2009) presented work on using context vectors to find NARROWING and BROADENING of senses over time by applying semantic density analysis. Each occurrence of a target word is mapped to its context vector, which follows the definition proposed by Schütze (1998). A context is considered to be 15 words before and after each target word. Two thousand words, the 50th to the 2049th most frequent word from the vocabulary are considered to be content-bearing terms C. Singular value decomposition is used to reduce the dimensionality to 100.

For a specific target word w, each occurrence of the word in the corpus can be mapped to a context vector. The SEMANTIC DENSITY of the word w in a specific corpus is defined as the average cosine similarity of the context vectors. A high similarity can be seen as a dense set of vectors and corresponds to words with a single, highly restrictive meaning. A low similarity is seen as a sparse set of vectors and corresponds to a word that is highly polysemous and appears in many different contexts. To reduce the computations, a Monte Carlo analysis was conducted to randomly choose n vectors for pairwise computation. To measure change in word senses over time, context vectors are created for a target word in different corpora (from different time points) and the semantic density is measured for each corpus. If the density of a word increases over time then it is concluded that the meanings of the word have become less restricted due to a broadening of the sense or an added sense. Decreased density over time corresponds to a narrowing of the sense or lost senses. Sagi et al. (2009) used four words in the evaluation that was conducted on the Helsinki Corpus (spanning texts from at least 1150–1710) divided into four sub-corpora; *do*, *dog*, *deer* and *science*. The first two were shown to broaden their senses, while *deer* was shown to narrow its sense. The word *science* was shown to appear during the period investigated and broaden its meaning shortly after being introduced.

Unlike in the work by Schütze (1998), the context vectors were not clustered to give more insight into the different senses. Instead, a random set of context vectors were selected to represent the overall behavior of a word. This means that even though there can be indication of semantic change there are no clues as to what has changed. What appears as broadening can in fact be a stable sense and an added sense. In addition, the method requires very balanced corpora, because the addition of attributes such as genre will affect the density.

3.1.2 Pointwise mutual information

Similar to the work described above, the work presented by Gulordava & Baroni (2011) builds on context vectors to identify semantic change over time. The authors used Google Books Ngram data, more specifically 2-grams (pairs of words) were chosen, so that the context of a word *w* is the other word in the 2-gram. Two separate sub-collections were chosen, the first one corresponding to the years 1960–1964 (the 60s) and the second one corresponding to 1995–1999 (the 90s). The content bearing words were chosen as the same for both collections and each count corresponds to the local mutual information similarity score. Two context vectors corresponding to the word *w* are compared by means of cosine similarity.

The assumption was that words with low similarity scores are likely to have undergone a semantic change, an assumption that was tested by manually evaluating a random sample of 100 words over all similarities. Five evaluators judged each of the words on a 4-point scale (from NO CHANGE to SIGNIFICANT CHANGE) based on their intuitions. The average value of these judgments was then used for each word and compared using the Pearson correlation measure. The results show that distributional similarity correlates the most with words that were more frequent in the 90s, while the frequency method correlates the most with words that were more frequent in the 60s. While this evaluation set is not freely available, it has been used by many others in follow-up work.

It is important to note that the evaluation measured the ability to detect not only change, but also to distinguish the degree of change. For better comparison with other surveyed methods, it would be useful to see how this method performs for the 100 most changed words, and as a comparison, to the 100 least changed words.

Rodda et al. (2016, 2017) present a method that relies on second-order similarities on the basis of positive pointwise mutual information scores, while Kahmann et al. (2017) propose using *context volatility* based on the significance values of a word's co-occurrence terms and their corresponding ranks over time. Three classes of change are evaluated on synthetic data while only one class, namely volatility, was evaluated on real data.

3.1.3 Temporal random indexing

Basile et al. (2016) presented one of few studies of the semantic change problem, before the LChange'19 workshop, in a language other than English. They focused on Italian and released a set of 40 words with their corresponding shifts in meaning. They made use of a word embedding method called TEMPORAL RANDOM INDEXING that builds on the authors' previous work (Basile et al. 2014). Each term gets a randomly assigned vector with two non-zero elements $\in \{-1, 0, 1\}$ and the assignments of all vectors are near-orthogonal. Then the corpus is split into sub-corpora where each one corresponds to a decade. The vocabulary in each sub-corpus is then modeled as the sum of all the random vectors assigned to each context word, normalized to downgrade the importance of the most frequent words.

The authors then used the change point-method proposed by Kulkarni et al. (2015) in two versions for detecting change, the pointwise change between two time adjacent vectors (*point*) (t_{i-1}, t_i) and the cumulative change (*cumulative*) between the sum of all vectors up to t_i and the vector for t_{i+1}.[18] An evaluation was performed manually. Given the set of change points returned by each method, the evaluation checked how many correct change points were detected among the top 10, top 100 and all of the returned change points. A change point is considered correct if it is found at the same time, or after the expected change point.

At the top 10 and top 100, the accuracy of the random indexing method performed as well as the log frequency baseline, and both outperformed other compared methods. The authors presented a time-aware evaluation as well as evaluated with which time delay the change points were found. The temporal indexing with *point* that got the best top 10 and overall scores had a time delay of, on average, 38 years with a standard deviation of 35. The best results were obtained by the random indexing and the cumulative method that, on average, had a delay of 17±15 and 19±20 respectively, however, with an accuracy of 12–16% on the detected change points.

3.1.4 Entropy

Tang et al. (2013) presented a framework that relies on time series modeling of the changes in a word's contextual entropy. For each period, a word was modeled as a distribution over its strongest noun associations. We can view this procedure as analog to first calculating a co-occurrence matrix and then performing dimensionality reduction, but here the dimension is reduced directly by associating w only to one noun from each context. The authors claimed that this helps represent different senses, as nouns have a high differentiating value. A WORD STATUS for w at time t is then the probability of these contextual nouns. To create a time

[18] Note that this is different from Kulkarni et al. (2015) who compare to t_0 for all time points t_i.

series, the feature vectors are represented by their entropy. The authors model linguistic change as an S-shaped curve (Kroch 1989) and apply curve fitting on the time series of the word status entropy to detect patterns for different kinds of change.

The authors used the Chinese newspaper *People's Daily*, spanning 1946–2004. The experiments show that the entropy time series of a word's feature vector can be used to identify different kinds of change, for example broadening by means of metaphorical and metonymic change. However, the values used for the classification are observations from the training data (all words that are classified also contributed to the finding of the thresholds). The experiment does not show the discriminating power of the variables on previously unseen data, a problem addressed and eliminated in the follow-up work.

Tang et al. (2016) attempted to cluster the contexts to find senses, and to classify the senses into different change types using the DBSCAN algorithm. The resulting clusters were considered synsets and their number reduced using the cluster's diachronic span and density.[19] The authors concluded that while it is possible to distinguish the different classes for each synset, the variables of the S-shaped curve were not sufficient for accurate classification. One important weakness is that the model only allows for one change event per word or sense (one S-curve). It is, however, possible that more than one change event occurs for each sense. In addition, the sense induction procedure was not evaluated properly; a different induction method (i.e., a different grouping of nouns into synsets) might provide better results.

3.1.5 Summary on co-occurrence-based methods

The co-occurrence based methods gave us a good starting point, and led the path into large-scale investigation of word-level sense change. Their greatest strength is offered by the interpretability of the vector spaces they create. However, they have come to be outperformed by the (static) embedding based methods surveyed next, primarily because the latter showed better performance when modeling semantics and detecting change.

3.2 Static neural embeddings

From 2014 and onwards, the largest body of work makes use of (neural) WORD EMBEDDINGS of different kinds. These embeddings are in many ways similar to

[19] This approach is considered sense-differentiated but we discuss it here since the description of the main algorithm is discussed here.

the co-occurrence based methods in that they create an n-dimensional vector space in which each word lives. Words close in the space should be similar in meaning. Static embedding vectors represent an average of the word over the whole corpus; the word *rock* is closer to *music* than *stone* in most modern corpora. However, unlike the count-based co-occurrence methods, embeddings rely on predicting rather than counting. Implicitly, they capture similar information, and have in some cases been shown to be equivalent mathematically, but in general they are better at abstracting and summarizing information from the corpus. In the same way as SVD vectors, the dimensions of the embedding vectors are not interpretable; they do not correspond to other words. Often, the closest words in the vector space are used to describe the meaning of the target word.

With a few exceptions, embeddings are individually trained on different time-sliced corpora and compared over time. This means that each representation for a word at different points in time lives in a different space, as a result of, among other things, the random factors. All different embeddings for a word must first be projected onto the same space before comparison. A few different methods have been used for projection. First, vectors are trained for the first time period t_1 independently of any other information. The follow up vectors for time $t_i, \forall i > 1$ are initialized with the vectors for time t_{i-1}. What happens in the case of words that are present in t_i but not in any time point before, is generally not specified, so the same initialization can be assumed as at time t_1 (see e.g., Kim et al. 2014 for more details). The second method projects words to a specified time period, typically the last one, using a linear mapping (see e.g., Kulkarni et al. 2015, Hamilton, Leskovec, et al. 2016 for more details and examples). Finally, two methods avoid mapping of vectors by comparing second order similarity vectors (see Eger & Mehler 2016) and a corpus trick for training time-specific vectors while utilizing the whole corpus at once (Dubossarsky et al. 2019). All of the papers in this section consider time series data and make use of different methods to detect changes compared to the average, first or last time period.

3.2.1 Initializing using previous time period

Kim et al. (2014) were the first to use neural embeddings to capture a word's meaning for semantic change detection. They used the Skip-Gram model (Mikolov, Chen, et al. 2013) trained on the Google Books Ngrams (5-gram) English fiction corpus. They created a neural language model for each year (with 200 dimensions), with the vectors being initialized by the vectors from the previous year. The years 1850–1899 were used as an initialization period and the focus for the study was 1900–2009. Vectors were compared over time using their cosine similarity. The 10 least similar terms (those believed to have changed their meaning the most) and the 10 most similar terms (stable terms), as outputted by the system, were inspected. The three closest neighboring words from 1900 and 2009 were used for verification of change.

Two words were investigated in more detail with respect to the time series of their cosine similarities; the difference in cosine similarity between the year *y* and 1900 was plotted against a time axis. This was compared to the average cosine similarity of all words as a baseline. It was clear that *cell* and *gay* deviated significantly from the average plot while the two stable terms *by* and *then* were more stable than the average. The comparison to the average of all words is an alternative method to comparing to negative words. This controls for the fact that not all words behave the same way as the changing ones and thus confirms the correct hypothesis. An alternative method is to compare, not to the overall average, but to the average of words in the same frequency span as the word under investigation (like in Jatowt et al. 2018). Comparing to other words in the same frequency span is important as there is evidence that very frequent words behave differently from very infrequent terms, in terms of semantic change (Hamilton, Leskovec, et al. 2016, Pagel et al. 2007, Lieberman et al. 2007).

In addition, the authors further grounded their results by investigating *n*-grams that contained the evaluated word from 1900 and 2009. We note that this, backwards referral to the original texts that contribute to a statistical hypothesis, is an extremely important step that is often overlooked by others.

The authors concluded that a word that has lost in popularity over time, and hence is not frequently used in succeeding time spans, will not update its vector and, therefore, change cannot be detected. They suggest combining embedding signals with frequency signals to detect such cases. No explicit evaluation with respect to outside ground truth was made, nor were the words marked for being correct or incorrect.

3.2.2 Change point detection

Kulkarni et al. (2015) presented an investigation of different statistical properties and their capacity to reflect statistically significant semantic change. Two questions were asked; how statistically significant is the shift in usage of a word over time? and at what point did the shift take place? Two things seem to be implicit in these questions. First, a shift in the dominant sense of a word (e.g., one existing, dominant sense handing over to another existing sense) was also considered a semantic shift. And secondly, a word has only one semantic shift. The authors noted that while many semantic changes occur gradually, there is a time when one sense overtakes the others and they considered this to be the change point, on lines of explanatory power of a sense (see Section 6.2.1).

For each word in a vocabulary, a time series was constructed over the entire time period. Each point t in the time series results in statistical information derived from the word's frequency, its part-of-speech distribution, or its semantic vector, corresponding to the usage in the sub-corpus derived at t, namely C_t.

A set of words were investigated in more detail and it was found that the distributional method performed better for some (a set of 11 words including *recording, gay, tape, bitch, honey*), while the syntactic method that utilizes part-of-speech information performed better for others (e.g., *windows, bush, apple, click*). A synthetic evaluation was presented, in which 20 duplicate copies of Wikipedia were used and the contexts of the words were changed artificially proportionally to a probability p_r. The larger the proportion of p_r, the better did both the distributional and the frequency method perform, with the distributional method outperforming the frequency method for all values of p_r. When the target and the replacement words were no longer required to belong to the same part of speech, the distributional method was outperformed by the syntactic method for low values of p_r.

The second evaluation was performed on a reference set of 20 words, compiled from other papers. Out of 200 words evaluated per method, 40% of the words from the reference set were found for the distributional method and 15% for the syntactic method. This was to some extent an experiment to capture recall of known changes. Finally, in the human evaluation, the top 20 words from each method were evaluated by three annotators. An interesting question arose when the time series of the syntactic and distributional methods were created. For both, the data at time t_i were compared to $t_0; \forall i$, where t_0 corresponds to the earliest possible time point and might have low quality due to sparse data and a high error rate. Would the method perform better if the information at t_i were compared to t_N where N was the last time point, to t_{i-1}, to an average of all time points, or to a joint modeling of all information at once?

Following Kulkarni et al. (2015), and drawing on the diverse information captured by the distributional models and frequency changes, two follow-up studies have attempted combining both, Stewart et al. (2017) and Englhardt et al. (2020).

3.2.3 PPMI-based compared to SGNS

Hamilton, Leskovec, et al. (2016) presented an evaluation of different embedding techniques for detecting semantic changes, both a priori known changes and those detected by the different methods. They evaluated six different datasets, covering four languages and two centuries of data.

The first embedding method was based on the positive pointwise mutual information score (PPMI). The second was a singular value decomposition (SVD) reduction of the PPMI matrix, often referred to as $\mathrm{SVD}_{\mathrm{PPMI}}$ in other work, and the third embedding method was the skip-gram with negative sampling (SGNS) (Mikolov, Sutskever, et al. 2013). The SVD and SGNS embeddings were aligned over time using the orthogonal Procrustes.

Four different tasks were evaluated: synchronic accuracy, detection of known pairs of change on both COHA and ENGALL, Google Books Ngram all genres, and discovery of new words that have changed on ENG fiction. The pairwise task considers the cosine similarity of a pair of words at each time slice, and correlates the value against time using Spearman correlation. For the detection of known pairs, a set of nine terms were compared with respect to a validation word. As an example, the term *nice* should move closer to *pleasant* and *lovely* and away from *refined* and *dainty*, resulting in four pairs.

Raw PPMI seemed to perform the worst while SVD performed better on smaller datasets, and SGNS better on larger datasets. The key novelty of this paper was the use of orthogonal Procrustes to align vector spaces, a method that has since been extensively used.

3.2.4 Summary on static neural embeddings

The static embedding methods paved the way for large-scale investigation of lexical semantic change and drew interest to the field from a sizeable portion of the NLP community. Their strength is that they model word meaning given the corpus on which they are trained, and do not rely on large pretrained models. Compared to the co-occurrence based models, they are also more effective at capturing word meaning.

Their downside is multi-fold: (1) they require a large number of words per time slice to create stable vector spaces, which means that they might be less applicable to languages with little digitized historical text; (2) when trained on individual datasets (corresponding to time-slices) they need to be aligned, a procedure that often introduces noise; and (3) they model the average meaning of each word on the basis of usage in the corpus, and hence do not allow for sense differentiation. Nonetheless, in SemEval-2020 Task 1, the majority of the methods were based on static embeddings, and dominated in both tasks compared to the deep contextualized embeddings; see Schlechtweg et al. (2020) for more details.

3.3 Dynamic word embeddings

A few distinct methods exist for creating dynamic word embeddings. Common to all of them is that they share some data across all time periods and that the resulting embeddings originate in the same space for all time periods. This reduces the need to align the vectors trained on separate time slices. However, each method uses different embedding techniques. It shows that, regardless of method for creating individual embeddings, sharing data across time is highly beneficial and can help reduce the requirements for large datasets (which are rarely available for historical, textual corpora).

3.3.1 Dynamic probabilistic skip-gram

The paper by Bamler & Mandt (2017) was the first of three to propose using dynamic word embeddings trained jointly over all time periods. The advantage of the method is two-fold. First, there is no need to align embedding spaces which can introduce noise (Dubossarsky et al. 2015, 2019), and second, the model utilizes information from all time periods to produce better embeddings and reduce the data requirements.

The authors proposed a Bayesian version of the skip-gram model (Barkan 2017) with a latent time series as prior. Their method is most like that of Kim et al. (2014), but information is shared across all (or all previous) time points. The priors are learned using two approximate inference algorithms, either as a FILTERING, where only past information is used (for time t_i all information from t_0 to t_{i-1} is used), or as a SMOOTHING, where information about all documents (regardless of time) is used. The resulting dynamic word embeddings can fit to data as long as the whole set of documents is large enough, even if the amount of data in one individual time point is small.

The authors compare their methods, dynamic skip-gram with filtering (DSG-f) and smoothing (DSG-s) with the non-Bayesian skip-gram model with the transformations proposed by Hamilton, Leskovec, et al. (2016) (SGI) and the pre-initialization proposed by Kim et al. (2014) (SGP).

The quantitative experiments aimed to investigate the smoothness of the embeddings between different, adjacent time periods. The experiments showed that joint training over all time periods is beneficial when training vectors for individual time periods, in the sense that the vectors do not move too radically from one year to another. Note, however, that this is a requirement included in the algorithm as part of the training.

The second set of experiments aimed at showing the capability of detecting semantic change. Again, the DSG-f and DSG-s outperformed SGI and SGP for the two smaller datasets, where the latter two methods had difficulty fitting a vector space to small amounts of data. For Google Books, the dynamic embeddings performed better in the sense that they were smoother.[20]

3.3.2 Dynamic PPMI embeddings

Yao et al. (2018) presented a second approach, with a different take on the word embeddings. Their embedding method relies on a positive pointwise mutual information matrix (PPMI) for each time period, which is learned using a joint optimization problem. In other words, embeddings for each time period were not first learned, then aligned, but rather learned while aligning.

The authors proposed these dynamic embeddings for both the semantic change problem and the diachronic word replacement problem. They investigated both problems using qualitative and quantitative evaluation. The authors crawled roughly 100k articles from the *New York Times*, published 1990–2016, together with metadata such as section labels. Four words were evaluated manually. The first two clearly illustrate the difficulty with modeling a word's meaning using a single representation: according to the model, *apple* has nothing to do with fruit from 2005 and onward and, since 1998, *amazon* is not a river.[21]

For the automatic evaluation, the authors automatically created a ground truth dataset using the section category of the 11 most discriminative categories from the New York Times. The comparison is done against three baselines, Static-Word2Vec (Sw2v, Mikolov, Sutskever, et al. 2013), Transformed-Word2Vec (Tw2v, Kulkarni et al. 2015) and Aligned-Word2Vec (Aw2v, Hamilton, Leskovec, et al. 2016). Both NMI and F-measures showed that the dynamic embeddings were better than the baselines, and while Sw2v and Aw2v followed closely, Tw2v showed a larger drop in performance. The authors suggest that this happened because local alignment around a small set of stable words was insufficient. While this seems reasonable, it does not explain why the Sw2v method (without alignment) performs better than the Aw2v method for all values of k for the NMI measure and was worse only for $k = 10$ for the F-measure.

[20] There are no precision values in the paper; the interpretation of the change results is left to the reader.

[21] In this dataset, the confusion of *apple* as a fruit and *Apple* as a company could be a consequence of the case normalization preprocessing step. The case of *Amazon* is different, since the mythological female warrior of antiquity, the jungle and the company are all proper nouns. It might also be a consequence of a change in the dominant sense of the words, to 'name of a company', and the representation method that might have difficulty capturing both at once.

Since information regarding most of the vocabulary is shared across time slices, the dynamic PPMI embedding method is considered robust against data sparsity, however, the authors did not mention any size requirements.[22]

3.3.3 Dynamic exponential family embeddings

A third method for creating dynamic embeddings was presented by Rudolph & Blei (2018). This method makes use of exponential family embeddings as a basis for the embeddings, as well as a latent variable with a Gaussian random walk drift. The key is to share the context vectors across all time points, but the embedding vectors only within a time slice. The results were compared to the results presented by Hamilton, Leskovec, et al. (2016) and the exponential family embedding (the static version).

As with Bamler & Mandt (2017), the dynamic embeddings performed better on unseen data. In a qualitative setting, a set of six example words were used to illustrate semantic drift, where the meaning of a word can change; its dominant sense can change; or its related subject matters can change. The authors presented the 16 words with the highest drift values for the U.S. senate speeches, and discussed a few of them in detail. They did however not present their view of these 16 words, or if any were considered incorrect. A change point analysis was presented, and contrary to Kulkarni et al. (2015), the authors did not make an assumption of a single change point, but no change point evaluation was presented.

A novelty presented by Rudolph & Blei (2018) is the investigation into the distribution of those words that changed the most in a given year. It does give some account of where interesting things happen to the language as a whole, and the authors recognize that the largest change occurred around the end of World War II (1946–1947). Another interesting spike occurred in 2008–2009 and what seems as the 1850s but these were not discussed further. The authors conclude by noting that the closest neighboring words over time show the semantic drift of words and can be helpful to discover concept changes.

There was no explicit differentiation between the change types. Instead, the absolute drift was computed as the Euclidean distance between the first and the last time points. Note that if the curve of changes in the embeddings behaves like a sine curve, there can be little difference between the first and the last change point, and the word can still experience substantial semantic drift in between.

[22] Yao et al. (2018) do not refer to the work of Bamler & Mandt (2017), and despite the different publication years, the work of Yao et al. (2018) was submitted before the work of Bamler & Mandt (2017) was published. Nonetheless, there is much overlap in the idea of jointly learning and aligning temporal vectors to produce smoother vector series for individual words.

3.3.4 Modeling change using a continuous time variable

Thus far, all of the dynamic methods have used static time bins despite the sharing of information across time. Different sizes of time bins result in different scales and granularity of change. Rosenfeld & Erk (2018) propose a method based on a modified deep SGNS architecture to model time as a continuous variable. This allows the method to capture gradual shifts and side-step having to make an a priori decision about bin size. The output is a differentiable function that given a time period *t*, and a word *w* (or *c* when the word is a context word), returns a time specific embedding for *w*. The method produces a static embedding for each time point, and one static time-independent word embedding (different for target and context words). Finally, there is a function for combining the word-independent time embedding with the time-independent word embedding using a linear layer. While the positive samples for each word are triples of the kind (*w*, *c*, *t*), the negative samples are chosen from the entire corpus and are time-independent. The negative examples are thus averaged across the entire dataset. The method is evaluated using 5 illustrative examples, and 45 synthetic change words with an automatic comparison to synthetic change rate.

The method is a sort of dynamic embedding in that it shares information across time, and avoids alignment while maintaining the possibility to create time-specific representations of individual words. However, there is a single time vector that influences all words, thus possibly limiting the method's capacity to model semantic change. The method is promising, in particular the functionality to model time as a continuous variable without the need to fix time bins beforehand (and the need for retraining if the decision changes), and an extension that allows multiple time vectors for different classes of words would be valuable.

3.3.5 Temporal referencing

The final method that falls into this category is based on a re-labeling trick to achieve the same goals of the other dynamic methods while using a static embedding method. Dubossarsky et al. (2019) train embeddings on a corpus as a whole, while relabeling target words during training with their time information, following Ferrari et al. (2017), Fišer & Ljubešić (2018), and Schlechtweg et al. (2019). A word *w* in a sentence c_1, c_2, w, c_3, c_4 from time *t* would be relabeled as c_1, c_2, w_t, c_3, c_4 only when *w* is a target word. This results in individual time-dependent embeddings for each target word but avoids alignment since they are all situated in the same space. The context embeddings are average embeddings across the whole corpus and thus suffer from bias towards time periods with more data. To avoid this bias in the evaluation, the authors train their embeddings on the decades between 1920–1970 of COHA, where data sizes are roughly equivalent. Two main methods are used, one based on SGNS and one on PPMI. Each of the models is trained in two flavors, one with independent time bins where the vectors are subsequently aligned ($\text{SGNS}_{\text{align}}$ and $\text{PPMI}_{\text{align}}$), and one on the temporal referenced corpus (SGNS_{TR} and PPMI_{TR}).

The authors begin by training all the four variants on shuffled corpora (where the information in each time bin is equally spread across the entire corpus while maintaining frequency properties (see Dubossarsky et al. 2017), and comparing how much semantic change remains. Semantic change is measured by average cosine distance (acd) for the whole vocabulary. SGNS_{TR} captures the largest true signal of change (i.e., the largest acd) measured as the difference between signal in the genuine and the shuffled corpora.

Next, the authors evaluated on a synthetic change task by taking multiple samples from the modern COCA, shuffled to mimic a synchronic language use. Pairs of words, for example *apple* and *pear* were used as donor and recipient, and over time, more and more of the contexts of the donor were inserted and relabeled with the recipient. Half of the 356 donor and recipient word pairs were chosen from SimLex-999 (Hill et al. 2015) to be related while the other half were chosen to be unrelated (a case that should be easier to detect because the words in the pair have widely different contexts).[23] In addition, an equal number of control words were chosen, where the increase in frequency was matched, but no synthetic change was introduced. For stable words, other w_{t_i} were consistently the closest neighbors, while for changing words other words were closer during periods of change. Finally, the word sense change testset (WSCT; Tahmasebi & Risse 2017) was used to evaluate the model's ability to detect semantic change in COHA. In both the synthetic, and the smaller real WSCT, the SGNS_{TR} outperformed the other models in terms of differentiating between changing and stable words.

While the method employed a static embedding approach, it benefitted from the sharing of contextual information, and the increase in corpus size when considering the corpus as a whole. Temporal referencing of corpora could in theory be used with any other static embedding method. It remains to compare the results to dynamic methods. The method has, however, been evaluated for both tasks in the SemEval-2020 Task 1, (Zhou & Li 2020), and performed well (3rd and 2nd in rank for the two sub-tasks).

[23] Additional ways of creating synthetic change types can be found in the work by Shoemark et al. (2019) and Schlechtweg & Schulte im Walde (2020).

3.3.6 Summary on dynamic embeddings

A first study to compare the dynamic embeddings proposed by Rudolph & Blei (2018), Bamler & Mandt (2017) and the static embeddings of Kim et al. (2014) was presented by Montariol & Allauzen (2019). The study shows that on low-volume data, dynamic models are better at detecting directed drifts. The base assumption that most of these models make is that most words do not change their meaning over time, and therefore, the context words can share one representation over time. If this holds true, the sharing of the majority of the text is highly beneficial and eases the case for languages that have fewer digitized historical words, but also enables the study of dialects or social groups where the amount of text is limited. The down-side of sharing context information across all time periods, can be the risk of not forgetting, that means, context words can contribute also in time periods where the association between the context word and the target word is weak or non-existent.

The development and in-depth study of further dynamic models has slowed in the past two years, probably as a result of the huge interest by the community in pre-trained, contextualized embedding methods.

3.4 Laws of sense change

Several authors have investigated general laws of sense change on the basis of large corpora. Here we summarize these laws.

Xu & Kemp (2015) evaluated two laws against each other, with respect to synonyms and antonyms. Using normalized co-occurrence vectors and the Jensen-Shannon divergence, Xu & Kemp (2015) investigated the degree of change for a given word measured as the difference in overlap between its closest 100 neighbors from the first and the last year of the Google Books Ngrams corpus. Using a set of synonyms and antonyms and a set of control pairs, the authors showed that, on average, the control pairs moved further apart than the synonyms and antonyms. They call this the LAW OF PARALLEL CHANGE: words that are semantically linked, like synonyms or antonyms, experience similar change over time and thus stay closer together than random words.

Dubossarsky et al. (2015) investigated the relation between a word's role in its semantic neighborhood and the degree of meaning change. Words are represented using their Word2Vec vectors trained on a yearly sub-corpus and similarity is measured using cosine similarity. Each yearly semantic space is clustered using *k*-means clustering (this can be seen as word sense induction but without the possibility for a word to participate in multiple clusters). A word's PROTOTYPICALITY (centrality) is measured as its distance to its cluster centroid (either a mathematical centroid, or the word closest to the centroid). Change is measured as the difference in cosine similarity for a word's vector in adjacent years, where the vector of the previous year is used as an initialization for the next, as in the work of Kim et al. (2014). The correlation between a word's centrality and its change compared to the next decade is measured. The 7,000 most frequent words in the 2nd version of the Google Books Ngrams English fiction corpus were investigated.

The authors showed that there is a correlation between a word's distance from the centroid and the degree of meaning change in the following decade. The correlation is higher for the mathematically derived centroid, compared to the word closest to the centroid. This indicates that the abstract notion of a concept might not necessarily be present as a word in the lexicon. Also the number of clusters play a role. In this study, the optimal number of clusters was 3,500, but this should reasonably change with the size of the lexicon. The trend was shown for a large set of words (7,000) over a century of data. This is the LAW OF PROTOTYPICALITY.

Hamilton, Leskovec, et al. (2016) suggested two laws of change, the LAW OF CONFORMITY, which states that frequently used words change at slower rates, and the LAW OF INNOVATION, which states that polysemous words change at faster rates. Polysemy is captured by the local clustering coefficient for a word in the PPMI matrix, which captures how many of a word's neighbors are also connected as a proxy for the number of different contexts that a word appears in.

At the same conference as Hamilton, Leskovec, et al. (2016), Eger & Mehler (2016) presented the LAW OF LINEAR SEMANTIC DECAY which states that semantic self-similarity decays linearly over time. They also presented the LAW OF DIFFERENTIATION, which shows that word pairs that move apart in semantic space can be found using the linear decay coefficient.

A follow-up evaluation presented by Dubossarsky et al. (2017) points to the need of proper control conditions when evaluating large-scale laws of change; the details of this study are further discussed in Section 6.4.

Without explicitly referring to laws of change, Ryskina et al. (2020) investigate the effects of semantic density and frequency on neologisms. By sampling 1,000 words that are novel in COCA compared to COHA, and 1,000 words that behave as their counterpart (controlling for frequency), they are able to show that neologisms are likely to appear in (a) semantic neighborhoods that grow fast in frequency, and to a lesser extent, (b) sparser areas of semantic space.

Rodina et al. (2019), followed by Kutuzov (2020) investigate the case of evaluative adjectives (e.g., *terrific*, *awesome*) compared to non-evaluative adjectives for English, Norwegian and Russian, and are able to show that, contrary to common belief, evaluative adjectives change at the same pace as non-evaluative adjectives.

3.5 Related technologies

Mihalcea & Nastase (2012) investigated the effect of word usage change in terms of the inverse problem, that of identifying the epoch to which a target word occurrence belongs, using a classification task (word sense disambiguation) with three epochs.

Takamura et al. (2017) targeted a slightly different but related task: to identify the difference in meaning between Japanese loanwords and their English counterparts. The authors recognized that semantic change in this context could mean that the Japanese loanword only adopted a single sense from a word's senses. Beinborn & Choenni (2019) go beyond one pair of languages and study semantic drift in multilingual representations.

A method to go beyond pure vector changes and look at the surrounding words is proposed by van Aggelen et al. (2016). They linked embeddings to WordNet to allow quantitative exploration of language change, for example, to which degree the words of a specific part-of-speech change over time. Concept change is studied via fully connected graphs by Recchia et al. (2016). Costin-Gabriel & Rebedea (2014) made use of the visual trends (using PCA) on Google Ngram viewer of words belonging to three classes: neologisms, archaisms and common words. Also Tjong Kim Sang (2016) made use of frequencies to detect neologisms and archaisms, using two measures. The first measured a delta of the last (known) and the first (known) relative frequency of a word, and the second measure checked the correlation between the relative frequency of a word to its average frequency. Both measures produced good, and complementary results, in manual evaluation. Morsy & Karypis (2016) framed their work in document retrieval and document similarity across time, and made use of link information and frequency information to implicitly account for language change. Azarbonyad et al. (2017) offered an alternative to change detection over time, and also studied detection of synchronic variation over viewpoints, similar to the work of Schlechtweg et al. (2019) that studied language change across domains.

Fišer & Ljubešić (2018) also study synchronic variation for modern Slovene lemmas under the assumption that the social-media texts would be "early adopters" of incipient semantic changes.

4 Sense-differentiated change detection

The methods presented in Section 3 do not currently allow us to recover the senses and therefore, little or no possibility of detecting *what* changed. Most methods show the most similar terms to the changing word to illustrate what happens. However, the most similar terms will only represent the dominant sense and not reflect changes among the other senses or capture stable parts of a word. In this section, we review methods that first partition the information concerning one word on the basis of sense information. There are several methods for detecting senses; some rely on WORD SENSE INDUCTION (also called DISCRIMINATION); some use topic models; and some rely on a general clustering mechanism.[24] A few of these attempt to track senses over multiple time spans. We will start by reviewing the topic-modeling and move to word sense induction methods. Finally, we will review the most recent methods based on deep contextual embeddings to detect sense change.

4.1 Topic-based models

Common to all topic-based models is that the topics are interpreted as senses. With the exception of Wijaya & Yeniterzi (2011) who partition topics, and Frermann & Lapata (2016) who use dynamic topic modeling, no alignment is made between topics to allow following diachronic progression of a sense. Topics are not in a one-to-one correspondence to word senses (Blei & Lafferty 2006, Wang & McCallum 2006) and hence newer induction methods aim at inferring sense and topic information jointly (Wang et al. 2015).

4.1.1 Detecting novel word senses

In their work, Lau et al. (2012) used topics to represent word senses and performed implicit word sense induction by means of LDA. In particular, a nonparametric topic model called HIERARCHICAL DIRICHLET PROCESS (Teh et al. 2004) was shown to provide the best results on the word sense induction task for the Semeval-2010 shared task. The number of topics was detected rather than predefined for each target word, which is beneficial when detecting word senses because all words have different numbers of senses in different datasets. The novel sense detection task was defined with the goal of detecting one or more senses assigned to a target word w in a modern corpus that are not assigned to w in an older reference corpus.

For each target word w, all contexts from both corpora are placed in one document D_w; the sentence with the target word, one sentence before and one after are used as a context. First, topic modeling was applied to the document D_w and

[24] The work of Tang et al. (2016) is presented in Section 3, under entropy-based methods as it is a follow up on Tang et al. (2013) where the entropy-based method is presented.

all topics were pooled (consisting of topics from both the modern and the reference corpora). Second, each instance of a target word *w* in the two corpora was assigned a topic. Finally, if a topic was assigned to word instances in the latter corpus but not in the former, then it was considered novel. A novelty score was proposed which considers the difference in probability for topic assignments normalized by a maximum likelihood estimate. The novelty score was high if the sense was more likely in the modern corpus and relatively unlikely in the reference corpus.

In the work by Lau et al. (2012), the written parts of the BNC reference corpus were chosen as the reference corpus, and the second, modern corpus was a random sample of the 2007 ukwac Web corpus (Ferraresi et al. 2008). Ten words were chosen for a more detailed examination, half of which were manually assessed to have experienced change while the other half had remained stable over the investigated time span. When ranked according to the novelty score, the five words with novel senses (henceforth novel words) were ranked in positions 1, 2, 3, 5, and 8. When repeating the experiment with frequency ratios, the novel words were ranked in positions 1, 2, 6, 9, and 10, indicating that pure frequency is a worse indicator than the novelty score in the case of two corpora that are wide apart in time and content.

In follow-up work, Cook et al. (2013) proposed a relevance score that incorporates a set of topically relevant keywords for expected topics of the novel senses, with the main aim of improving the filtering of non-relevant novel senses. In this work, two sub-corpora of the GIGAWORD corpus for the years 1995 and 2008 are used. The experiments in Cook et al. (2013) differ from that of Lau et al. (2012), in that instead of using a pre-defined set of evaluation words, Cook et al. (2013) used the top 10 words of the novelty score, the rank sum score, and a random selection for further investigation. The evaluation was conducted in a lexicography setting by a professional lexicographer. Half of the words found using the novelty score had no change in usage or sense. From the words found using the rank sum scores, all words were of interest. From the randomly chosen words only three words were of interest. The interesting cases were then analyzed by a lexicographer and found to belong to two different classes; having a novel sense (4 plus one of the randomly chosen ones) or in need of a tweak/broadening (9 plus two of the random ones).

A more extensive evaluation was performed by Cook et al. (2014) where two corpus pairs were used, the BNC/ukwac and the SiBol/Port corpora (that consists of a set of British newspapers, similar in theme and topics, from 1993 and 2010), with 7 and 13 words with novel senses respectively, and a significantly larger set of distractors.

Though it was not suggested by the authors in this series of papers (Lau et al. 2012, Cook et al. 2013, 2014), the method could be used to find the inverse of novelty as well. If a topic is assigned to instances in the reference corpus but not in the second corpus, then the sense can be considered outdated or, at least, dormant. Overall, the method proposes the use of topic modeling for word sense induction and a simple method for detecting novel senses in two separate corpora, both by using novelty scores and by incorporating topical knowledge. The senses were, however, not tracked; the exact same sense is expected to be found in both the reference and the modern corpus. Assume for example that there is a sense s_i in the reference corpus that does not have a match in the modern corpus, and a sense s_j that has a match in the modern but not in the reference corpus. If s_i is similar to s_j, then the two senses could be linked, and possibly considered broadening or narrowing of each other. The difference in s_i and s_j could also be a consequence of random noise. By not considering the linking of topics, and only two time points, the complexity was significantly reduced. Drawing on work like that proposed by Mei & Zhai (2005), it remains for future work to track the topics over multiple time periods so additional change types can be detected beyond novel senses.

4.1.2 Clustering and tracking topics

The work of Wijaya & Yeniterzi (2011) addressed some of the weaknesses of the novel sense detection methods, by targeting automatic tracking of word senses over time, where word senses were derived using topic modeling.

The experiments were conducted on Google Ngram data where 5-grams were chosen in such a way that the target term w was the middle (third) word.[25] A document D_w^i was created for each year i consisting of all 5-grams where w was the third word. Then these documents were clustered using two different methods. The first experiment made use of the k-means clustering algorithm and the second experiment made use of the TOPIC-OVER-TIME ALGORITHM (Wang & McCallum 2006), an LDA-like topic model. In the *k-means* experiment, topics were considered to have changed if two consecutive years were assigned to different clusters.

For the topic-over-time clustering, two topics were created and the algorithm outputs a temporal distribution for each topic. At each time point, there was only one document. While not directly specified, the strength of a topic i, for $i = 1, 2$ for a time period was likely the assignment of topic i to the document at time j.

[25] The authors do not specify the time span of the data, and consequently we estimate it to be roughly 500 years, that of the Google Ngram dataset.

When the most probable topic for a document changes, so does the sense of the word target word w.[26]

A few different words are analyzed; two words changed their dominant sense, *gay* and *awful*. Two words added a sense without losing their previously dominant sense, *mouse* and *king*, where the latter also became a reference to Martin Luther King. In addition, the authors tested the method for changes to a named entity, Iran's change from monarchy to republic, and John F. Kennedy's and Bill Clinton's transitions from senator to president. Both algorithms captured the time of change, either by a change in cluster or topic distribution. Two change classes are used for the analysis but the algorithm does not differentiate the different kinds.

Adjectives do not seem well suited for the method as their meaning was not well captured by topic models. This might be because topic modeling is not optimal for capturing word senses (Boyd-Graber et al. 2007). In general, the work presented by Wijaya & Yeniterzi (2011) was preliminary but it was the first paper to provide an automatic method for working with more than one sense of a word to find out *what* happened in addition to *when*. There was no proper comparison between the different algorithms to indicate which method performs better or to quantify the results. Two questions remain unanswered. One is, how many of the 20 clusters in k are reasonable? Another is, how often, on average, do we see a change in cluster allocation for the k-means clustering? Nevertheless, the overall methodology of using clustering to associate different topics or documents with each other could be a promising direction.

4.1.3 Dynamic topic models

Frermann & Lapata (2016) proposed a dynamic topic model, called SCAN, that differs from the above in several aspects. First, the topic models in their proposal are not independently created for each period, but rely on the adjacent time period. Implicitly, there is a tracking of senses over multiple time periods. Second, each topic can exhibit change over time, to capture subtle changes within a sense. Like the topic-over-time algorithm, this dynamic Bayesian model produces a set of fixed topics with a time distribution to show their probability over time. It also allows for changes over time within each topic. An example was given to

[26] Using the k-means algorithm on documents does not represent a fully sense-differentiated method. The topic-over-time method represents only two senses active at the same time, and those are constant over time. These two senses correspond to having one representation for two different major senses over different times, where one hands over to the other. Still, we have chosen to categorize the method among the sense-differentiated methods.

highlight the importance of allowing sense representations to change. The word *mouse* changed in one of its senses, from the 1970s, where words like *cable, ball, mousepad* were important, to *optical, laser, usb* which are important today. All the while both representations stood for the computer device sense of *mouse.*

The DATE corpus, spanning the period 1700–2010, was used for the experiments. The corpus was tokenized, lemmatized and part-of-speech tagged, and stopwords and function words were removed. All contexts around a target word w from a year t were placed in one document, and a time span was 20 years. A context window of size ±5 was used, resulting in what can be seen as an 11-gram with the target word in the middle, as the 6th word. For two out of three experiments, the number of senses was set to 8. In the third experiment, the number of senses was set to 4.

The first experiment was a careful analysis of four positive words, namely *band, power, transport* and *bank.* For each word and topic number (1 … 8), there were (at most) 16 different topical representations, one per time period. On average, 1.2 words were exchanged, a number that was controlled by the precision parameter. No quantification of this number (in relation to the precision parameter, or on its own) was given. The words that stayed among the top-10 did, however, move in rank over time, which signified change without the words being exchanged.

The second experiment considered novel sense detection (Lau et al. 2012) and borrowed its evaluation technique from Mitra et al. (2015) and its relevance ranking from Cook et al. (2014). The results for eight time pairs, with a reference and a target time, were presented. In this experiment, the number of senses was set to 4. As a baseline, the same model was used to learn topics independently (i.e., without the dependency on previous time periods) and was called SCAN-NOT. For this, the topics were matched across time periods using the Jensen-Shannon divergence measure. The topics with the lowest Jensen-Shannon divergence were assigned to the same topic number. There was no lower threshold so topics that were very different, but still had the lowest divergence could be assigned to the same topic number. Novelty scores were calculated using the relevance score to determine when a topic represents a novel sense. A total of 200 words were identified as sense birth candidates. For the 8 time pairs, SCAN performed better than SCAN-NOT in 6 cases.

The final experiment[27] related to word meaning change and made use of the test set presented by Gulordava & Baroni (2011). The test set consists of 100 words

[27] The authors presented a fourth experiment on the SemEval-2015 DTE task for identifying when a piece of text was written, which we have not presented here.

annotated on a 4-point scale, from no change to significant change. The novelty score (as defined by Cook et al. 2014) was calculated on the same 100 words, comparing the 1960s with the 1990s, and 8 senses per word. The result was the Spearman's rank correlation between the novelty scores and the human ratings from the test set. The correlation score for SCAN was 0.377, as compared to 0.386 reported by Gulordava & Baroni (2011) on a different, and larger training set. The SCAN-NOT (0.255) and frequency baseline (0.325) performed worse than SCAN.

The study leaves open questions. For example, the authors did not properly argue for the choice of 8 topics per word, and from the experiments it seems like a large number; for the word *power* three senses were identified; 'institutional power', 'mental power' and 'power as supply of energy'. These were distributed over 4, 3 and 1 topics, respectively. What would happen with a lower number of topics? The time span of 311 years was partitioned into 8 time periods, which significantly reduced complexity of evaluation. How the method performs with smaller time spans and more time periods remains to be evaluated.

While novelty of senses was evaluated in detail, there was no discussion of how to differentiate change types or how the method would perform on control words. For the small, in-depth evaluation presented on four words, we saw that all 8 associated topics change[28] over time for each word. For example, the 'river bank' sense of *bank* should reasonably exhibit a stable behavior, not change so radically over time, to allow the distinction of a stable sense from a changing sense. The evaluation of change in individual topics also remains for future work. Is the change in top-10 words or the change in probability of the same set of words over time reasonable for a sense?

The SCAN-method represents an interesting approach that contains most of the necessary components for studying semantic change. Topics were modeled (for individual time periods but with a dependence on previous times) and automatically linked over time, and were themselves allowed gradual change. This could enable tracking of individual senses for a word and their rise and fall; it could link them according to concepts and separate the stable senses from the changing ones. We highly encourage additional studies exploring these possibilities. An extension to SCAN is seen in the GASC model (Perrone et al. 2019) where also genre information is incorporated. This is shown to be particularly useful for Ancient Greek, where the lack of data and the long diachronic time span make it harder to find semantic change in a reliable way.

[28] Change was measured in terms of topical strength, the overlap of the top-10 words between adjacent time periods was not specified.

4.1.4 Summary on topic-based models

Topics offer easy and robust modeling and division of words into semantic areas. Though there is no proven direct link between topics and senses, there is a division based on topical usage as evident in the text. Compared to the WSI-based methods surveyed next, these models offer higher recall, as a word will always belong to a topic. Compared to the static neural embeddings, they also offer the possibility to reproduce which sentences contributed to a specific topic, at different points in time. This enables close reading and evaluation of the models which is of high interest for studies in, e.g., conceptual history, history, and lexicography.

While the study into static neural embeddings has evaluated several methods for aligning vectors from different independent spaces as a way of tracking vectors, the study of lexical semantic change using topic models has limited itself to either using the same topics over time, or a Kim et al. (2014)-like model where one model is initialized with the information from the previous one. A more thorough investigation of different kinds of models for tracking and the effects of these on the change detection seems like a natural and important next step.

4.2 WSI-based models

Models based on WORD SENSE INDUCTION (WSI) were utilized by Mitra et al. (2014, 2015), Tahmasebi (2013), and Tahmasebi & Risse (2017) to reveal complex relations between a word's senses by (a) modeling senses per se using WSI; and (b) aligning senses over time. The models allow us to identify individual senses at different periods in time and Tahmasebi & Risse (2017) also merge senses into linguistically motivated clusters.

4.2.1 Chinese whispers

The work of Mitra et al. (2014) was followed up by Mitra et al. (2015), which presented a more comprehensive analysis. In this review, we will refer to the 2015 work, which almost completely includes the earlier work.

The aim of the experiments was to track senses over time and to identify if the sense changes were due to birth (novel senses), death (disappearing senses), join (broadening of senses by two senses joining into one), and split (narrowing of a sense by a sense splitting into two). The core part of an approach like this is the method for finding sense clusters. In this work, the method used for detecting senses was the Chinese whispers algorithm (Biemann 2006). It is based on clustering a co-occurrence graph. For each word, a set of 1,000 features are kept, where features are derived from bigram relations. A pair of words are linked in the graph if they share a sufficient number of features. The local graph is clustered by starting with all nodes as individual clusters and then merged in a non-deterministic manner, to maximize edge weights of the clusters. To overcome some of the randomness, the procedure is run multiple times and the results are pooled.

Once the clusters are in place, the tracking begins. For each two adjacent time periods, the set of clusters for a word w are compared and the word overlap between any two clusters is measured. To detect birth or join, the overlap is divided by the size of the cluster in the newer period and, inversely, the older period for death and split. A set of criteria determine to which class the clusters belong.

Two datasets were used in the experiments, Google Books Ngrams (1520–2008) and Twitter (2012–2013). The former dataset was split into eight periods where the first spans 1520–1908 and the last spans 2006–2008. The aim was to have roughly equal amounts of text in each time span. The clustering was applied in each time period separately, and compared to all subsequent time periods (and between Google Ngram and Twitter for a cross-media analysis). A set of candidate births (ranged from roughly 400 to 4200) were detected between each time span. These changes are considered stable if, for example, a novel sense s that was detected in t_2 compared to t_1 was also novel in t_3 compared to t_1.

The evaluation was performed using two methods, one manual and one automatic. For the manual evaluation, the time period 1909–1953 is compared to 2002–2005. A set of 48 random birth words and 21 random split/join words were inspected manually. The accuracy was 60% for birth cases and 57% for split/join. A set of 50 births were evaluated with respect to Twitter and Google Ngrams, out of which 70% were correct (between datasets no joins or splits were found).

The automatic evaluation is done with respect to WordNet where clusters for a word w are mapped to a synset of w. The method makes use of a synchronic sense repository for detecting sense changes. The mapping is done on the basis of the words in each cluster and their presence as synset members. Roughly half of the clusters are mapped to a synset, but no formal evaluation is conducted. A birth is a success if a cluster s_{new} gets assigned a WordNet synset ID that is not assigned to any of the word's clusters in the earlier period. A split is a success if the two clusters in the new time period have different synset IDs ($s_{\text{new1}} \neq s_{\text{new2}}$) and one of them is the same as the old cluster ($s_{\text{new}\, i} = s_{\text{old}}$, for $i = 1, 2$). The join success criteria are analogous to the split criteria, where the new and old time period have swapped places. For the manual evaluation, the period 1909–1953 was compared to all succeeding periods. While average accuracy scores were not given, the histogram showed values ranging from roughly 48% to 62% for births, from 38% to 53% for splits, and from 30% to 64% for joins.

The method does not track senses over multiple time periods; the tracking is done pairwise. This means that the functionality is currently not in place to track a novel sense that is later joined with another. While there is a filtering that requires that a novel sense should still be novel in the next time period, the tracking is not done over the entire time period.

4.2.2 Curvature clustering

The work of Tahmasebi (2013) and Tahmasebi & Risse (2017) has a long-standing basis in (manual) studies related to diachronic conceptual change on the basis of the curvature clustering algorithm. The aim is to track word sense clusters over time, for individual senses of each word, and to group senses into semantically coherent clusters. Related senses should be grouped together, while unrelated senses should be kept apart.

The basis of this line of study is the word sense clusters, which rely on the curvature clustering algorithm (Dorow et al. 2005) applicable to nouns and noun phrases that appear in coordination. Dorow et al. (2005) investigated the quality of the clusters on WordNet for modern data (British National Corpus) and Tahmasebi et al. (2013) evaluated the quality with respect to historical data. The quality of the clusters remained high despite the higher number of OCR errors, but the number of extracted clusters dropped with higher error rates. The experiments were conducted on the (London) Times Archive and the New York Times annotated corpus, on yearly sub-corpora. The resulting dataset spanned 1785–1985 and 1987–2007.

The cluster sets for a target word *w* were compared over subsequent years. The comparison was done using a modified Jaccard similarity (to boost similarity between clusters of largely different sizes but with high overlaps) and a WordNet-based similarity measure based on the Lin (1998) measure. In the first phase, clusters that were similar enough to be considered the same over time (including some random noise) were grouped. These groupings correspond to stable senses over an arbitrary time span. In the next phase, these groupings were compared across all time periods. This two-step procedure was used to reduce the complexity, as otherwise the possible transitions between clusters grow exponentially with the number of clusters and time periods. After these two first steps, there were a set of linked senses over time for a target word. As a final step, the individually linked senses were grouped into semantically coherent groups, while unrelated senses belonged to different groups.

The method allows for the detection of broadening and narrowing, splitting and merging of senses, novel related and novel unrelated (e.g., neologisms) senses, and stable senses. Each change event was monitored individually, hence a word could first have a novel sense that later changed by means of, for example, broadening. These were then considered two separate change events. The stable senses could belong to two different categories, those words that had no change events and were stable over the entire time span, and those that experienced change in another sense. An example of the first category is the word *horse* and of the latter category is the word *rock*, where the ‘stone’ sense is stable while the ‘music’ sense is first added (as a novel unrelated sense as it is not related to any previously existing sense), and later changed by means of broadening.

The test set consisted of 35 change events corresponding to 23 words, and 26 non-change events. Eleven of these corresponded to stable words without other change events and the remainder corresponded to words that had change events related to their other senses. In addition, the authors also evaluated the time delay with which the change was found with respect to both a ground truth dataset and to the first available cluster representing a given sense or change. On average, 95% of all senses and changes were found among the clusters, showing the upper limit for the word sense induction method on the dataset. Eighty-four percent of the change events could be discriminated and correctly identified. Only related, novel senses could not be found properly, most likely due to little or no word overlap in the contexts.

The average time delay was presented as a time span between two time points. The first represents the manually chosen outside world (and can be the time of invention or the first attested use of a word sense) but does not need to be valid for this specific dataset. The second represents the time the (automatic) word sense induction method can detect evidence of a sense or change. If the gap between these two time points is large, there is either little evidence in the datasets, or the WSI method was unable to detect the sense. The true time delay lies between these two points. For detected senses and changes, the time delay is on average 6.3–28.7 years. For the change events that can be discriminated and correctly identified, the time delay is slightly higher, 9.9–32.2. In particular, existing senses of words with change events have a time delay of 11.7–59.0, while the corresponding number for words without change events is much lower, 2.7–20.5 years. These delays can be compared to those present presented by Basile et al. (2016) who found, on average, a time delay of 38 years for change in the dominant sense. This speaks to the fact that words are unlikely to change their meaning if they are used frequently.

The strength of the method is the possibility of tracking senses on an individual basis; and to allow for certain parts of a word to stay stable while other parts change independently of each other. The ‘food’ sense of an *apple* does not disappear because ‘the company Apple’ is the more popularly used sense. All senses are tracked over each year, which increases the complexity but keeps a fairly high granularity for change detection. The authors did not filter any results and hence presented no precision.

4.2.3 Summary on WSI-based methods

The WSI-based methods are the only ones where the unsupervised outcome has been evaluated with respect to word senses per se. This has the advantage of offering a higher certainty on what exactly is modeled and tracked, however, at least the curvature clustering algorithm offers low coverage of senses. Models based on deep contextualized embeddings offer similar functionality, where individual representations for each word usage exist, and these need to be grouped in such a way that the groups represent senses. Once grouped into senses, methodology for tracking of the senses can be drawn from the WSI-based methods.

4.3 Deep contextualized embeddings

Among the sense-differentiated methods, a few make use of DEEP CONTEXTUALIZED WORD EMBEDDINGS, typically pre-trained BERT embeddings (Devlin et al. 2019). For contextual representation of a token w, information from an entire context, for example the sentence in which w participates, is used to deduce the token representation.

The first work to employ BERT is presented by Hu et al. (2019). The sense-differentiation is made using the *Oxford English Dictionary* (OED). Each target word w is looked up in OED and for each sense s_i of the target word, up to ten example sentences are extracted. The sense embedding for s_i is computed as the average of the token representations for the sentences corresponding to the sense. This procedure allows also outdated senses, as these are present in the sense repository of OED, however, the sense representations are static and calculated in advance.

The authors begin by evaluating the method’s capability to accurately describe senses. They sample sentences for a set of ambiguous words corresponding to one of the word’s senses s_i. Next, the contextual embedding of each word is attained. This embedding is compared to the sense embeddings found for all of the senses of w. If the closest sense, in terms of cosine similarity, belongs to the correct sense s_i, this is considered a success. The accuracy of the method is above 92% and thus the authors conclude that the method is suitable for representing senses given a sense repository with sample sentences.

A lemmatized version of COHA, divided into decade bins, is used for the experiments. Words that appear a minimum of 10 times per year in at least 50 consecutive years are considered as target words. Using the derived sense embeddings for each target word, the proportion of each sense is calculated over time. The authors follow Tang et al. (2016) and perform a smoothing by decomposing the diachronic sense proportions into a trend component and a noise component using polynomial curve fitting, and use the trend component for further analysis.

The method is evaluated for its capacity to detect sense change by using the manually annotated dataset presented by Gulordava & Baroni (2011), and improve on both the original study, and the results reported by Frermann & Lapata (2016). The authors continue to study SENSE COMPETITION (where the dominant sense changes over time) and COOPERATION (where several senses follow the same trajectory). Among 3,220 studied words, almost 23% of the studied words undergo a change of dominant sense, at least once, that is not as a result of change in part-of-speech. For sense cooperation, senses should be similar or related in meaning (a high cosine similarity), in addition to following the same trajectory, and together overtake the dominant sense. Over 31% of the changes fall in this category, and shows that the study of multiple senses and their interaction has a high impact for change detection.

The method above can be seen as a semi-supervised approach given that the OED is used to guide the definition of senses. In the past year, several groups have attempted unsupervised methods where clustering is used to find sense or usage clusters.

Giulianelli et al. (2020) make use of the base-uncased version of BERT without any fine-tuning, and create context representations for each occurrence of a word in each time period (decadal periods from COHA 1910–2009). They follow Schütze (1998) and use k-means clustering, and go further by searching for the best solution for k ranging over 2–10. The clustering is a global clustering in that they cluster the set of all usages across time, a cluster then contains usages from all time periods where the "sense" was valid. They compare clusters from different time periods using Entropy difference and Jensen-Shannon divergence. Following Sagi et al. (2011), the authors also calculate an average pairwise distance between all context representations from two different time periods, disregarding cluster information. They evaluate using 16 of the words introduced by Gulordava & Baroni (2011) with added annotation for usage types across 20 year periods, spanning in total 100 years. The authors provide an interesting analysis of the results and qualitatively analyze which kinds of change can be found.

Also Martinc, Montariol, et al. (2020) cluster contextual representations derived from BERT, after first having fine-tuned BERT on COHA. They find that affinity propagation provides better results than k-means for different values of k on the full 100 words introduced by Gulordava & Baroni (2011).

4.3.1 Summary on deep contextualized embeddings

So far, the majority of the work on using contextualized embeddings has focused on BERT, a pre-trained model that can be fine-tuned on the corpus under study. The strength of these methods is to some extent similar to the ones of static embeddings, they have a high coverage. An advantage, however, is that because they are vectors in a joint space, they do not need aligning. They also offer easy comparison (for example by means of cosine similarity). However, while pre-trained models are very robust, having been trained on billions of tokens, they can be dominated by information that does not stem from the corpus under investigation. Which means, they can model meaning primarily from the corpora used for training, e.g., primarily model American English compared to Singaporean English. If the corpus under study is small, fine-tuning might not alleviate this problem. A method like ELMo, which is also contextualized, but lightweight enough to be trained on the corpus itself, can be beneficial in certain cases (Kutuzov 2020).

In SemEval-2020 Task 1 (Schlechtweg et al. 2020), many deep contextualized models were compared to methods relying on static embeddings. The competition showed that static embeddings outperform the contextualized ones, under the settings in the task. One property that most likely had an influence is the lemmatization of the corpora. It remains as a task for future work to compare these methods on non-lemmatized corpora as well. In addition, it remains to investigate how sentence-based embeddings can be best grouped to represent word senses.

4.4 Aligned corpora

The work conducted by Bamman & Crane (2011) sought to track the rise and fall of Latin word senses over 2000 years. Adopting an old idea (Dagan et al. 1991, Dagan & Itai 1994), they used two aligned corpora in different languages for translation of words to help approximate the senses of the word. The number of different translations in language B will provide a probable guess on how many

Table 1.4: Datasets used for diachronic conceptual change detection. Non-English *

Sagi et al. (2009)	Helsinki corpus
Gulordava & Baroni (2011)	Google Ngram
Wijaya & Yeniterzi (2011)	Google Ngram
Lau et al. (2012)	British National Corpus (BNC), ukwac
Cook et al. (2013)	Gigawords corpus
Cook et al. (2014)	BNC, ukwac, Sibol/Port
Mihalcea & Nastase (2012)	Google books
Basile et al. (2016)	Google Ngram (Italian)
Tang et al. (2013, 2016)*	Chinese People's Daily
Kim et al. (2014)	Google Ngram
Kulkarni et al. (2015)	Google Ngram, Twitter, Amazon movie reviews
Mitra et al. (2015)	Google Ngram, Twitter
Hamilton, Leskovec, et al. (2016)	COHA, Google Ngram
Eger & Mehler (2016)*	COHA, Süddeutsche Zeitung, PL[a]
Azarbonyad et al. (2017)	New York Times Annotated Corpus, Hansard
Rodda et al. (2017)*	Thesaurus Linguae Graecae
Frermann & Lapata (2016)	DATE corpus
Takamura et al. (2017)	Wikipedia (English and Japanese)
Kahmann et al. (2017)	Guardian (non-public)
Tahmasebi & Risse (2017)	Times Archive, New York Times Annotated Corpus
Bamler & Mandt (2017)	Google Books Ngrams, State of the Union addresses, Twitter
Yao et al. (2018)	New York Times (non-public)
Rudolph & Blei (2018)	ACM abstracts, ML papers ArXiv, U.S. Senate speech
Rosenfeld & Erk (2018)	Google Ngram (Eng. fiction)
Hu et al. (2019)	COHA
Dubossarsky et al. (2019)	COHA
Giulianelli et al. (2020)	COHA

[a] Patrologiae cursus completus: Series latina.

different senses are valid for the word in language A. The translation mechanism also helps to determine the frequency with which the instances of the target word are assigned to the senses; the more often the target word is translated to word *i* in language B, the more often the sense *i* is assigned to the target word in language A.

The results clearly showed that sense variations could be measured over time and pointed to a change in the predominant sense over time for five chosen terms. The method is far more beneficial for studying words and their meanings over time than studies based on word frequency. However, it is limited as it requires a translated corpus to train the word sense disambiguation classifier. In addition, it does not allow the senses to be aligned over time to follow the evolution of senses and their relations.

4.5 Comparison

Finally, Table 1.4 gives an overview of the datasets used, and Table 1.5 provides a summary with respect to the most important aspects and differences of the studies reviewed in this section.

5 Computational modeling of diachronic word replacement

While diachronic conceptual change, including semantic change, corresponds to the semasiological view, diachronic word replacement corresponds to the onomasiological view. These can be seen as two sides of the same coin, and resolving diachronic lexical replacement is a prerequisite to be able to completely handle diachronic conceptual change.

Several works have attempted to characterize diachronic replacement and the processes that are governing it. Pagel et al. (2007) proposed a general hypothesis that nouns are replaced more easily than verbs as well as that frequent words undergo less replacement. Others claimed that rich synonym networks speed up replacement (Vejdemo & Hörberg 2016). More recently Karjus et al. (2020) have demonstrated that the change in communicative needs of speakers and the competition related to topic salience can help to explain the lexical replacement process in languages.

Ullmann (1959) already discussed taxonomies of types of lexical replacement processes from a theoretical and conceptual point of view. Influenced by the computational approaches and actual applications, we roughly distinguish here the following types of diachronic replacement:

Table 1.5: Comparison of methods for diachronic conceptual change detection. (m) illustrates that the proportion of positive and negative change words is unspecified. Experiments on automatic or manual analysis are divided by | . Values separated by / represent results on different datasets or algorithms. Time spans are given in years, unless specified with *m* for months. A: automatic eval method; B: broadening; M: manual eval. method; N: narrowing; P: pair entity; S: single entity.

	prechosen		top	entity	eval. method	time		# classes:	modes	
	# pos	# neg				span	# points	classes	time | sense	
									aware | diff	
Sagi et al. (2009)	4	0		S	M	569	4	2: B/N	no	no
Gulordava & Baroni (2011)	0	0	100 (m)	S	M	40	2	1: change	no	no
Tang et al. (2013)	33	12		S	M	59	59	3: B/N/novel/change	no	no
Kim et al. (2014)	0	0	10+10	S/P	M	110	110	1: change	yes	no
Kulkarni et al. (2015)	20	0	20/20/20	S	M/A	105/12/2	21/13/24	1: change	yes	no
Hamilton, Leskovec, et al. (2016)	28	0	10/10/10	S/P	M	200/190	20	1: change	no	no
Rodda et al. (2017)	0	0	50	S	M	1200	2	1: change	no	no
Eger & Mehler (2016)	0	0	11+10	S/P	M	200/190	20/19	1: change	no	no
Basile et al. (2016)	40	0		S	M	170	17	1: change	yes	no
Azarbonyad et al. (2017)	24	0	5+5	S	M	20/11	2/2	1: change	no	no
Takamura et al. (2017)	10	0	100+20	S/P	M	-	2	1: change	no	no
Kahmann et al. (2017)	4	0		S	M	11*m*	48	1: change	no	no
Bamler & Mandt (2017)	6	0	10	S/P	M	209/230/7	209/230/21	1: change	no	no
Yao et al. (2018)	4|1888 (m)	0		S	M/A	27	27	1: change	no	no
Rudolph & Blei (2018)	5	0	16	S	M	9/64/151	9/64/76	1: changed	no	no
Rosenfeld & Erk (2018)	5|45	0		S/P	M/A	110	continuous	1: change	no	no
Dubossarsky et al. (2019)	13|356	19|356		S	M/A	50	5	2: related/unrelated	yes	no

Table 1.5 (cont.)

	prechosen		top	entity	eval. method	time		# classes:	modes	
	# pos	# neg				span	# points	classes	time	sense
									aware	diff
Wijaya & Yeniterzi (2011)	4	2		S	M	500	500	2: change/novel	yes	yes
Lau et al. (2012)	5	5		S	M	43 y	2	1: novel	no	yes
Cook et al. (2013)	0	0	30	S	M	14	2	1: novel	no	yes
Cook et al. (2014)	7/13	50/164		S	M	43y/17y	2/2	1: novel	no	yes
Mitra et al. (2015)	0	0	69/50	S	M/A	488/2	8/2	3: split/join/novel	no	yes
Frermann & Lapata (2016)	4	0	200	S	M/A	311	16	2: change/novel	no	yes
Tang et al. (2016)	197	0		S	M	59	59	6: B/N/novel/change	no	yes
Tahmasebi & Risse (2017)	35	25		S	M	222y	221	4: novel/B/N/stable	yes	yes
Hu et al. (2019)	100	0	10	S	A	200	20	2: dom. sense ch./sense coop.	no	yes
Giulianelli et al. (2020)	16 (m)	0	0	S	M/A	100	10	1: change	no	yes

(1) LEXICAL REPLACEMENT relates to words of any part of speech and its detection requires sense information. Words may have different sets of senses at different times and some of the senses can be replaced by others. Examples include *foolish* that replaced *nice* for the 'foolish' sense of the latter, and *cool* that replaced *relaxed*.[29]

(2) Terms that describe the same entity/object at different times and represent different names of that entity/object. For example, *Myanmar* has replaced *Burma* as the name of the country, and both refer to the same object (same identity). Note that an object here needs to be a named entity (i.e., it refers directly; see Section 2.1). Furthermore, multiple names can be used to refer to the same object at the same time, and some names can substitute for others over time. The latter represents the phenomenon of diachronic NAMED ENTITY CHANGE.

(3) Terms that are instances of the same type that were valid at different times, for example, the names of US presidents. Note that the instances could be exclusive at any given time point (i.e., there is only one US president at a given time point). Here, the analogy consists in the fact that the instances are of the same type/concept and not influenced by the attributional similarity of the instances (e.g., whether president George W. Bush was really similar in his character or other attributes to president Bill Clinton).

(4) The last type is TEMPORAL ANALOGS, which are terms similar due to shared role, attributes, functions despite time gap, yet they do not belong to the other three types. Analogy in general is a cognitive process of transferring information or meaning from a particular subject called the analog or source to another subject called the target. Temporal analogy could be considered a subtype of analogy because it is a comparison of two subjects that existed at different times based on their similarity or equivalence. One reason for finding analogous terms at different times is providing support for querying document archives. For example, *Walkman*, *Discman* and *iPod* could be considered analogs as portable music devices existing at different times.

The three latter types (without lexical replacement) are conceptually depicted in Figure 1.1. Most of the previous work originates from the information retrieval

[29] The latter replacement is seen as a synchronic variation as both words, *cool* and *relaxed*, are used in different populations to mean the same thing. In the former case, *nice* has completely lost its 'foolish' sense.

field aiming at finding diachronic replacements for a given target query word. From a practical viewpoint, the ability to find diachronic replacements could have many applications ranging from uses as components in larger systems such as search engines (e.g., for query suggestion), knowledge graph maintenance, educational applications, or, in NLP pipelines and in broad uses aiming at comprehensive text understanding as well as commonsense reasoning. Below we survey a number of works on automatically finding diachronic replacements over time. However, we note that most of them do not use the sense information of a word, hence effectively treating a word as having one sense (i.e., often its dominant sense). We mainly focus on works related to finding replacement types (3) and (4), i.e., named entity replacements and temporal analogs.

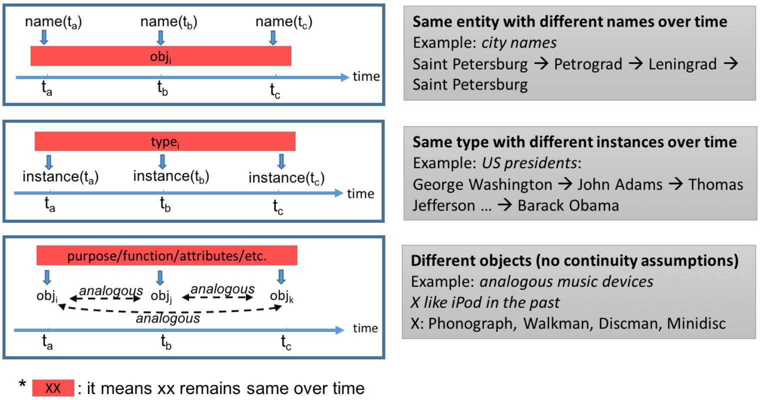

name(t_a)
name(t_b)
name(t_c)
obj_i
time
t_a
t_b
t_c
Same entity with different names over time
Example: city names
Saint Petersburg → Petrograd → Leningrad → Saint Petersburg
type_i
instance(t_a)
instance(t_b)
instance(t_c)
time
t_a
t_b
t_c
Same type with different instances over time
Example: US presidents:
George Washington → John Adams → Thomas Jefferson ... → Barack Obama
purpose/function/attributes/etc.
obj_i
analogous
obj_j
analogous
obj_k
analogous
time
t_a
t_b
t_c
Different objects (no continuity assumptions)
Example: analogous music devices
X like iPod in the past
X: Phonograph, Walkman, Discman, Minidisc
* XX : it means xx remains same over time

Figure 1.1: Conceptual view of three types ((2), (3) and (4) from the list above) of diachronic replacements.

Berberich et al. (2009) were probably the first to propose reformulating a query into terms used in the past in order to support user search within document archives spanning over long time periods. The task was defined as follows: given a query $q = q_1, q_2, .., q_m$ formulated using terminology valid at a reference time R, identify a query reformulation $q' = q'_1, q'_2, .., q'_m$ that paraphrases the same information need using terminology valid at a target time T. They measured the degree of relatedness between two terms used at different times through context comparison using co-occurrence statistics. A hidden Markov model was used for query reformulation; it considered three criteria of a good reformulation: SIMILARITY, COHERENCE, and POPULARITY. In particular, the similarity criterion requires that q_i and q'_i have high degree of across-time semantic similarity, while coherence means that q'_i and q'_{i-1} should co-occur frequently at time T to avoid combining unrelated terms. Finally, q_i' should occur frequently at time T to avoid unlikely query reformulations. This approach may require recomputation each time a query is submitted because it needs a target time point for the query reformulation.

Kaluarachchi et al. (2010) proposed that semantically identical words (or named entities) used at different time periods could be discovered using association rule mining to associate distinct entities with events. Sentences containing a subject, verb, and object are targeted and the verb is interpreted as an event. Two entities are then considered semantically related if their associated event is the same and the event occurs multiple times in a document archive. The temporally related term of a named entity is used for query translation (or reformulation) and results are retrieved appropriately with respect to specified time criteria.

Kanhabua & Nørvåg (2010) extracted time-based synonyms of named entities from link anchor texts in Wikipedia articles, using the full article history. Because of the limited time span of Wikipedia, they extended the discovered time of synonyms by using a burst detection method on the New York Times Annotated Corpus. Unfortunately, link information, such as anchor text, is rarely available and thus limits the method to hypertext collections. The authors evaluated the precision and recall of the time-based synonyms by measuring precision and recall in the search results rather than directly evaluating the quality of the synonyms found.

Tahmasebi et al. (2012) proposed a method called *NEER* for discovering different names for the same named entities (e.g., *Joseph Ratzinger* and *Pope Benedict XVI*, *Hillary Rodham* and *Hillary Clinton*). It relied first on detecting the periods that had a high likelihood of name changes and analyzed the contexts during the periods of change to find different temporal co-references of named entities. The key hypothesis was that this approach could capture both the old and the new co-reference in the same context. The underlying assumption was that named entity changes typically occur during a short time span due to special events (e.g., being elected pope, getting married or merging/splitting a company). Co-references were classified as direct and indirect. Direct co-references have some lexical overlap (e.g., *President Obama* and *Barack Obama*), while indirect ones lack any lexical overlap (e.g., *President* and *Barack Obama*). The proposed method first identified potential change periods via burst detection. Bursts related to an entity were found by retrieving all the documents in the corpus containing the query term, grouping them into monthly bins, and running the burst detection on the relative frequency of the documents in each bin. After NLP analysis, the method creates a co-occurrence graph of nouns, noun phrases and named entities from documents mentioning the input entity. The next step collapsed the co-references based on their lexical similarity and merged their contexts into co-reference classes. All terms in the context of a given co-reference class were then considered as candidate indirect co-references.

Tahmasebi et al. (2012) conducted experiments on the New York Times dataset using 16 distinct entities corresponding to 33 names and 86 co-references (44 indirect and 42 direct). Using a random forest classifier they achieved a precision of 90% on known time periods and 93% on found periods. The proposed method was later applied for query suggestion in search engines using temporal variants of a query (Holzmann et al. 2012) and for detecting named entity evolution in the blogosphere (Holzmann et al. 2015).

As is typical, there is low overlap between contexts of temporal analogs, solutions that rely on measuring context overlap do not work well. Distributed word representations (e.g., Mikolov, Sutskever, et al. 2013) can be useful for avoiding the problem of low context overlap. Given the representations trained on the distant time periods (typically, one derived from the present documents and another from documents published in the past), matching words across time could be done through transformation. This essentially means aligning relative positions of terms in the vector spaces of different time periods. Zhang, Jatowt, Bhowmick & Tanaka (2016) and later Szymanski (2017) used a linear transformation matrix for finding translations between word embeddings trained on non-consecutive time periods for detecting temporal analogs. The inherent problem in this kind of approach is the difficulty of finding a large enough training set, given the variety of domains, document genres, and arbitrary time periods for finding temporal analogs. A simple solution proposed by Zhang, Jatowt, Bhowmick & Tanaka (2016) assumes that frequent and common terms in both time periods can be easily acquired and used for optimizing the linear transformation matrix. This idea is based on the observation that most frequent words are known to change their semantics across time only to a small degree (Hamilton, Leskovec, et al. 2016, Pagel et al. 2007, Lieberman et al. 2007). Initializing word embeddings using embeddings trained on previous time periods (Kim et al. 2014) is difficult given the potentially long gaps between the two periods on which the vector spaces were trained. The potential lack of data from the intermediate periods can be another problem. The authors also successfully experimented with using terms that were computationally verified to have undergone little semantic variation across time as training instances for the transformation matrix. They did this by comparing sequentially trained word representations from consecutive time periods. Another improvement was the introduction of a local approach that relied on transforming automatically selected reference terms for a given query, which are supposed to ground the meaning of the query. Such transformed reference terms were then compared with the reference terms of candidate analogs, which had been generated by the previously described global transformation approach, with a linear transformation matrix. In other words, the global transformation approach was effectively extended with a method that locally constrains a query by transforming selected context terms called reference terms and then compares these terms with the ones of candidate analogs. The reference-to-reference term similarity measure relies not only on comparison of transformed vectors but also on comparison of transformed vector differences. The idea behind comparing vector differences was to capture the relation of a query (or a candidate analog) and its reference term. Three methods were suggested for proposing the reference terms from candidate context terms: PMI, clustering, and hyperonym detection using shallow processing (Ohshima & Tanaka 2010) in an attempt to reflect the relevance, diversity, and generality of the reference terms, respectively. Experiments were done on manually constructed ground truth data consisting of pairs of temporal analogs using precision at different cutoff points and mean reciprocal rank (MRR). The results showed that the local approach using reference terms selected from hyperonyms of a query (and of candidate terms) performed the best. The authors also demonstrated that correcting OCR errors by using a simple approach based on word embedding similarity and word frequency greatly enhances the quality of results.

More recently, Zhang et al. (2017) proposed using a set of transformation matrices based on different hierarchical clusters over the vocabularies in the two time periods. The thinking was that a single linear transformation matrix is insufficient for obtaining a good mapping between vector spaces of different periods. However, they found that using a series of matrices, such that each corresponded to a given hierarchical cluster of terms, and aggregating their results performed better.

Orlikowski et al. (2018) compared a number of models that rely on operations on word embeddings using nine different concepts on a corpus of Dutch newspapers from the 1950s and 1980s. Following Kenter et al. (2015), the authors assumed the notion of diachronic concept change involving the core concept terms and characterizing concept terms. Based on that model, the characterizing terms are expected to change over time, while the surface forms of the core terms are assumed to stay the same. The problem of concept change at a particular time point is then reduced to the problem of predicting valid characterizing terms for a core concept term.

All the approaches proposed so far have relied on sense-agnostic solutions, essentially, mixing all the senses (or relying on the dominant sense). A future improvement would be to move into the direction of finding analogous terms with respect to their senses or topics/aspects (sometimes called viewpoints). For the example of the latter, consider *Walkman* which corresponds to *iPod* due to their similar function as a 'music device', while *PC* can be a reasonable analog when regarding *iPod* as a 'game player'. A queried term, for example, an entity, may contain multiple aspects and the temporal analogs could be different depending on the particular topic/aspect. In this regard, Zhang et al. (2019) demonstrated a simple solution for an aspect-based temporal analog retrieval that takes additional terms as input to restrict the meaning of a user query to a particular viewpoint or aspect. The proposed solution also utilizes a neural network to realize non-linear term-to-term mapping.

Furthermore, all the approaches, with the exception of the work by Tahmasebi et al. (2012) and Kaluarachchi et al. (2011), need clearly specified time periods for comparison. While typically one of the periods represents the present (i.e., the time when a present-day user needs some information), the others can be any period in the past. It is, however, not always feasible to require users to specify specific periods for which temporal analogs need to be output. In many scenarios, it may be assumed that the user wants to know all the analogs from the past; hence, methods that can provide ranked results based on the agglomeration of results collected from different time periods should also be proposed.

Outputting evidence for automatic explanation of term similarity is a related problem to estimating similarity across time. The approach proposed by Zhang, Jatowt & Tanaka (2016) relies on providing evidence of terms' similarity over time by outputting explanatory context terms and then extracting sentences that reveal the shared aspects between temporal analogs. For example, for the input query pair `ipod` and `walkman`, the pairs of explanatory terms could be `music–music`, `device–device`, `apple–sony`, `mp3–cassette`, and so on. Note that the input is now the pair of query terms instead of a single term, as it is in the temporal analog retrieval task, and the output is the ranked list of term pairs. Term pairs are ranked based on their relevance to the input query pair as well as the intra-similarity between the pair elements and their relations to query terms (both similarities are computed after applying transformation).

For this, Duan et al. (2019) proposed an approach that uses joint integer linear programming and entity-oriented typicality analysis to generate multiple pairs of corresponding terms across time.

Turney & Mohammad (2019) used WordNet synsets and Google Books Ngram data to investigate the competition of words belonging to the same synset. The authors used a supervised learning approach (a naive Bayes classifier) to predict future leaders in the synset based on a range of features like word length, the characters in the word, and the historical frequencies of the word. The weaknesses of this approach are the assumption of the stability of synsets over time (the last two centuries) and inability to model words moving between synsets.

Finally, recent work by Karjus et al. (2020) demonstrated a simple way to find candidates of diachronic lexical replacements that is based on the comparison of word frequency changes and word semantics as represented by latent semantic analysis (LSA). Their competition model assumes that words which increase in frequency can substitute semantically similar words that experience decrease in their frequency around the same time.

6 Methodological issues and evaluation

6.1 Evaluation and hypothesis testing

Today, it is considered more or less *de rigueur* to accompany a proposed new method in computational linguistics with an AUTOMATIC, FORMAL, QUANTITATIVE EVALUATION. This reflects a healthy development towards greater objectivity in reporting results, but it also comes with a greater responsibility on the part of the researchers to ensure that the evaluation metrics provide a true measure of the accuracy of the proposed method.

Given the vast amount of digitized information now available to us, there is currently a unique possibility to develop and test methods for detecting language change. However, the amount of data limits the possibility to use expert help and manual efforts in the detection phase. It is also a limiting factor in the evaluation phase as there are to date only a few existing, open datasets for diachronic conceptual change that can be used for evaluation purposes.[30]

Specific to this problem is the grounding of diachronic conceptual change in a given corpus. When does a word appear for the first time with a new or changed sense in a given corpus? As a consequence, there are few automatic evaluation methods. Instead, there is a large variety of techniques, datasets and dimensions that are used in the existing literature. Most previous works have made use of manual evaluation while some have made use of WordNet for evaluation purposes. We argue that WordNet is not appropriate for evaluation for two main reasons. First, there is no indication in WordNet of *when* a word's meaning changed or a new sense was added. Second, when datasets span hundred years or more,

[30] The SemEval-2020 Task 1 on unsupervised lexical semantic change detection presents a first large multilingual resource. The organizers report over 1,000 annotation hours and close to 20,000 € in costs for a human-annotated dataset for two time periods.

WordNet does not sufficiently cover the vocabulary or word senses in the dataset. The same holds for Wikipedia, which often covers changes but lacks time information (Holzmann & Risse 2014). In addition to the lack of data and resources for evaluation, there are no evaluation methods or metrics that have themselves been properly evaluated.

Note that downstream applications, e.g., IR systems, can of course be evaluated in the normal way for such applications, which we will not describe here. Rather we will focus on methods for evaluating lexical change as uncovered by the methods surveyed here. A reasonable assumption would be that such an evaluation regime will also be useful – at least in part – for evaluating concrete downstream applications.

At least in the context of this literature survey, we would like to step back and see computational linguistics-style formal evaluation as part of a larger endeavor, as a central and necessary, but not sufficient, component of (LINGUISTIC) HYPOTHESIS TESTING. In particular, since the gold standard datasets which make up the backbone of our formal evaluation procedures are generally extremely expensive to create, there is an understandable tendency in our community to reuse existing gold standards to the greatest possible extent, or even re-purpose datasets originally constructed with other aims in mind.[31] However, such reuse may be in conflict with some assumptions crucial to the original purpose of the dataset, which in turn could influence the results of the evaluation.

There are two central (typically tacit) methodological assumptions – i.e. hypotheses – made in the work described in the previous sections, and especially in work on diachronic conceptual change detection and classification (Sections 3 and 4):

1. APPLICABILITY: the proposed method is suitable for uncovering diachronic conceptual change.

2. REPRESENTATIVENESS: the dataset on which the method is applied is suitable for uncovering diachronic conceptual change using this method.

Since most current approaches are data-driven – i.e. the data are an integral component of the method – these two factors, while logically distinct, are heavily interdependent and almost impossible to keep apart in practice, and we will discuss them jointly here.

With a few notable exceptions, to which we will return below, there is also often a third tacit assumption:

[31] Or even generate synthetic, simulated data assumed to faithfully reflect authentic data in all relevant aspects.

3. FALSIFIABILITY AND CONTROL CONDITIONS: positive evidence is sufficient to show 1 and 2.

Assumption 3 comes at least in part from the common practice of evaluating diachronic conceptual change using lists of attested such changes, and is often logically wrong.

We will now take a closer look at these assumptions.

6.2 Applicability and representativeness

The first major difficulty when evaluating the results of diachronic conceptual change is the *evaluation of a representation $r \in R$* of a meaning of a word w or a word sense s_w. When is r a correct and complete representation of w or s_w? Typically, this boils down to determining if a set of words, derived by clustering, topic modeling or from the closest words in a word space, indeed corresponds to the meaning of a word or word sense. In the case of multi-sense tracking, it is also important that the set of representations in R are a complete representation of w such that all its senses are represented in a correct way. The evaluation of individual word senses is analogous to the evaluation of word sense induction (see Agirre & Soroa 2007, Navigli 2012 for more details and an overview).

Another related, more subtle, source of methodological muddles may be a misunderstanding of what is being investigated. Liberman (2013) points out that the notion of "word" used in a paper by Petersen et al. (2012) is very far from how this term is understood by linguists, and the purported statistical laws of vocabulary development as evidenced in the Google Ngram dataset can be due to many other irrelevant factors, foremost of which is varying OCR quality, but also "tokenization" as a faithful model of wordhood (Dridan & Oepen 2012).

Linguists have long recognized that "language" is a nebulous term, at best designating a convenient abstraction of a complex reality. However, this does not mean that any language sample should be considered equally representative. Especially corpus linguists have spent much intellectual effort on the question how to compile representative language samples, where it is clear that "representative" generally must be interpreted in relation to a specific research question. We mention this here, since we feel that it is important to be clear about what the changing entity is when we investigate lexical change. Given that linguists generally consider speech to be the primary mode of linguistic communication, are we happy investigating mainly written language, following a long tradition of "written language bias" (Linell 2005/1982) of general and perhaps especially computational linguistics? Or given that the language should belong to every member of its speaker community, are we satisfied modeling the language of a select small social stratum (Henrich et al. 2010, Søgaard 2016)? An interesting aspect of this discussion comes into play when employing pre-trained models like BERT: do the results live in the dataset being studied, or stem from the pre-trained model? Does this mean that the results are more general, though we typically have little say in what data is used in the pre-training phase? Whatever the answers to these questions are, they need to be addressed. We should also recognize that to be able to use statistical inference from a corpus sample to the population as a whole, the sample must be random. Due to the above stated reasons, and many more, we cannot assume that written corpus data are ever a random sample of a language as a whole, and hence, we cannot use what we learn on a corpus to infer about the language in general (Koplenig 2016). To be able to reason about the language as a whole, we need many experiments from a wide range of sources to converge on the same conclusion.

The second major difficulty concerns the COMPARISON OF WORD SENSES (via their approximations) over time. Because the word senses are approximations derived from time sliced corpora, the representations at different time points can be different without there being any actual sense change. Two factors can play a role:

Factor 1: Imagine a set of contexts C that contain word w. If we split C into two random sets C_1 and C_2, such that $C_1 \cup C_2 = C$, the representations of w in C_1 and C_2 respectively will be different. Assuming that $|C_1|$ and $|C_2| \to \infty$ the difference in representation of w for C_1 and C_2 should go to 0. However, this is rarely the case, since our datasets are finite in size and we see a difference in representations. Because we often use single genres of data, novels, news papers etc., we are likely to enhance this randomness effect; if a word is not used in a certain context due to missing underlying events, then the word sense will not be present. By using a mixed set of sources, we could reduce this effect. We see the same effect for representations of a word w if C_1 and C_2 belong to two different time periods.

Now, if C_1 and C_2 represent two adjacent time periods, the task of diachronic conceptual change becomes to recognize how much of the difference in the representations of w that is due to this randomness effect and how much is due to actual semantic drift.

Factor 2: Imagine that the representation of w is a set of words $u_1, \ldots, u_n$ for time t_i and $v_1, \ldots, v_n$ for time t_j. If each v_j is a diachronic word replacement of u_j, then the entire representation of w can be replaced between t_i and t_j without there being any change to the sense of w. While it is unlikely that all words are replaced between any t_i and t_j, the risk of this effect increases the further apart the time periods.

In other words, in order to argue that some instance of lexical variation constitutes a case of diachronic conceptual change based on (massive) corpus evidence, it it generally not enough to ascertain that the variation correlates with different time slices of the dataset. It is also necessary to ensure that no other relevant variables are different between the time slices. The original Culturomics paper (Michel et al. 2011) has been criticized for not doing this, by Pechenick et al. (2015) and Koplenig (2017b), among others. This is also held forth as a strong point of the smaller COHA dataset by its creator (Davies 2012). This pitfall can be avoided by devising control conditions, but even so the purported diachronic effect may conceivably disappear for other reasons as well, e.g., if some other variable unintentionally correlates with time because of how the data were compiled. An interesting example of this is the fact that two random, trending variables will have a moderate to high correlation despite being completely random (Koplenig & Müller-Spitzer 2016). Here, the correlation stems from the fact that most diachronic corpora increase in volume over time and not necessarily from underlying semantic changes.

Another interesting reduction is the *n*-gram model, that automatically limits the amount of available information. To date, there has been little, if any, discussion in the diachronic conceptual change detection field to cover the effects of using *n*-grams rather than a full dataset with running text.[32] What happens when we remove words out of *n*-grams (which is the case when we only keep the *k*-most frequent words)? How many *n*-grams still have sufficient information left? What is the distribution of the remaining 1-, 2-, 3-, 4- and 5-grams after the filtering? This is particularly important when we consider those works that keep the *k*-most frequent words without normalizing over time, and hence have a modern bias among the kept words. If we start with equal samples over time, how many *n*-grams contribute over time?

An important aspect of representativeness is language coverage. While it is certainly true that the studies surveyed here are on a much larger scale than any historical linguistic studies heretofore conducted, it is nevertheless misleading to characterize traditional historical linguistic investigations as "based on small and anecdotal datasets" (Dubossarsky 2018: 2). This ignores the combined weight of the diversity of active observations painstakingly and diligently made over two centuries on many languages and language families by a large number of scholars highly trained in linguistic analysis, observations which are continually shared

[32] Gale et al. (1992: 233) note that in their experiments on word-sense disambiguation, they "have been able to measure information at extremely large distances (10,000 words away from the polysemous word in question), though obviously most of the useful information appears relatively near the polysemous word (e.g., within the first 100 words or so)."

and discussed in the professional literature of the discipline. Against this is set computational work on massive textual (published) datasets largely confined to one language – the norm – or a typologically and geographically skewed sample of a few languages. While such work undoubtedly will contribute valuable data points to our collective knowledge of lexical change, in order to make solid linguistic claims about this kind of language change, it would be desirable to conduct equivalent experiments on as many languages as possible (see e.g., Bender 2009, 2011, 2016).

6.2.1 Factors involved in evaluation of diachronic conceptual change detection

6.2.1.1 Granularity

The first and most important factor that impacts evaluation is to determine the granularity on which to evaluate. Typically, change is evaluated with respect to change in the dominant sense of a word. That is, changes are not evaluated individually for all the senses of a word; instead, meaning change is evaluated for the form (text word or lemma), i.e. mixing all its senses. Having a single representation per time period significantly reduces the complexity as it does not take into consideration what happens individually for each sense of a word. If a word has at most s ($s \in S$) senses per time period over t ($t \in T$) time periods, the number of unique senses is bound by $S \cdot |T|$. To compare all senses pairwise between time periods there are at most $|T|{\cdot}\mathrm{S}^2$ comparisons needed. If we wish to evaluate the *similarity graph* created by the senses in each time period, where edges correspond to similarity between two senses $s_i \in t_i$ and $s_j \in t_j$, there are $S^{|T|}$ possible paths. In comparison, for the single representation case, the number of unique senses are $|T|$ and the number of necessary comparisons is $|T|-1$ and there is only one path to evaluate. The number of time periods affects this complexity, and while some use yearly subcorpora, others use decades, reducing the time periods to compare by one order of magnitude.

6.2.1.2 Context

What is considered the *context* of a word differs largely between different works and is to some extent determined by the choice of dataset. A context ranges from 30 words surrounding w (Sagi et al. 2009) to the word before and after (Gulordava & Baroni 2011). When the Google N-gram data is used, the context can be at most a window of 5 words (from 4 words before or after, the word w being the first or last word, or 2 words before and after, the word w being the 3rd word). For the pre-trained contextual embeddings of BERT, the sentence before, the target sentence and the sentence after are used as a context. What information is used as a context affects the representation.

6.2.1.3 Words included in the evaluation

An important part of evaluation is to determine which words to evaluate. Here two methods are employed; a set of pre-determined words, or the (ranked) output of the investigated method or methods. The former has the advantage of requiring less effort and reduces the need to conduct a new evaluation for each new run, with e.g., new parameters. The downside is, however, that the evaluation does not allow for new, previously unseen examples. Please note that using only positive examples can result in false conclusions: if we assume that the method always concludes change and is tested only on words where we expect change, it will be 100% correct regardless of the choice of words. We believe that using negative examples to show the method's capacity to differentiate the positive and the negative examples is needed, and that the falsifiability assumption stated in Section 6.1 is generally wrong.

Pre-chosen testset

- positive examples (words known to have changed)
- negative examples (words known to be stable)

Output of algorithm

- on the basis of a pre-determined measure of change (e.g., largest or smallest cosine angle between two consecutive time periods)
- randomly chosen set of words

Most commonly, single words are used in evaluation, but it is becoming increasingly common to study the relation between (known) word pairs. That means, two words, typically one that is under investigation and one that represents the changed word sense, are evaluated with respect to their similarity over time. If a change takes place between the pair, this is used to confirm the hypothesis of diachronic conceptual change. Examples include (*gay, homosexual*) that become more similar over time, or (*gay, happy*) that become less similar over time. Both would confirm the same hypothesis about change in meaning for the word *gay*. Thus far, word pairs have always been used in a pre-chosen fashion. Choosing the word pairs that have the highest degree of change increases the computations by a polynomial factor. If we assume that there are n words at time t, and worst case, a new set of words for each time period, then there are $(n^2)^t$ pairs available. Typically, the situation would be much less extreme and only a fraction of the vocabulary is exchanged per time period (the more, the further apart the time periods are). Moreover, the reference term to be chosen for judging the changes of a target term should itself have stable meaning over time. For example, when tracking the similarity between *gay* and *happy* in order to detect or understand the sense change of the former, one implicitly assumes that *happy* does not undergo significant semantic change over the time period of comparison.

6.2.1.4 Evaluation technique

Evaluation can be conducted manually or automatically. Manual evaluation is done either with respect to intuition or pre-existing knowledge, or against one or more resources (dictionaries, encyclopedia etc.). Automatic evaluation is performed with respect to external resources, e.g., WordNet, or intrinsically where some evaluation metric is compared over time, e.g., statistically significant difference in the direction of the word vectors (Kulkarni et al. 2015).

Evaluation of temporal analog search often follows IR style evaluation settings. For a given query a ranked list of analog terms is presented and metrics like precision/recall (Tahmasebi et al. 2012) or precision@1, precision@5 and MRR (Zhang, Jatowt, Bhowmick & Tanaka 2016) are used based on the rate of correct analogs found in the top ranks.

6.2.1.5 Change types included in the evaluation

Evaluation for each word can be a binary decision; yes/no, there has been change, but it can also take the time dimension into consideration. The change is correct if it is found at the expected time point, or it is correct with a time delay that is measured. In addition to the binary decision, there are different change types (recall Table 1.2, page 16, for a list of change types considered in this literature). The more types are considered, the more complex the evaluation becomes. With one exception, different change types are considered only for sense-differentiated methods, while word level change groups all changes into one class. Typically, change means a shift in the dominant sense of a word. For example, Apple becomes a tech company and adds a dominant meaning to the word *Apple*. However, its ‘fruit’ sense is not gone but is very much still valid.[33] Still, the change in

[33] Note, however, that in written standard texts this “change” will partly be an artifact of preprocessing; lowercasing all text will increase the likelihood of conflating the common noun

dominant sense from ‘fruit/food’ to ‘technology’ is considered correct in a word level change setting.

6.2.1.6 Time dimension

The time span of the data makes a difference for evaluation. The further back in time, the harder it is to evaluate since there are fewer resources that cover the data (e.g., no reference resources such as dictionaries/wordnets/wikipedias for historical senses, etc.) and fewer experts to perform in-depth manual evaluation. The complexity is increased with the number of included time points. The more time points, the more complex the evaluation as there are more comparisons to evaluate.

The evaluation of time is an extremely complex matter; should it be done with respect to the outside world or the specific dataset under investigation? The complexity of the evaluation differs largely depending on the choice. To compare to the outside world means to make use of dictionaries and other knowledge sources to determine when a word came to existence, changed its meaning or added a sense (see Viola & Verheul 2020 for example). The resource or resources used for this determination need not be tied to the dataset used and there are regional variations in uptake of new politics, technology, culture, etc., that in turn affect language use. Newly coined terms, or senses can be due to an invention, one or a few influential sources, or an event and in such cases, be simpler to pinpoint in time. If the change, however, is due to a slow cultural shift or idiom that increases in popularity, it becomes very difficult to pinpoint the time of change. An analogy is that of fashion; when did the bob cut come into fashion? When the first ever person got such a haircut? Or the first celebrity showed it off on the red carpet (where is was better noticed and more likely to be duplicated)? Or when we can measure that a certain percentage of women had the hair cut as attested by e.g., school pictures or driver’s licenses? In manual attestation of diachronic conceptual change it is common to discuss the explanatory power of a sense in a given time; however, that is hard to translate into a specific time point. A more or less arbitrary threshold can be used to translate an increasing (or decreasing) curve into a binary yes or no that can be used to specify a time point.

If we wish to evaluate with respect to the dataset, there is an added difficulty compared to the above. If the word itself is not novel, then it requires word sense disambiguation to find the first occurrence of a new or changed sense; when was a word used in a specific sense for the first time in the dataset? If existing sense

apple and the proper noun *Apple*. It is also in fact likely that the “dominant” sense of *apple* is an artifact of the dominant modality and genre, and not a fact of language repositories are not available, the senses must first be induced and then assigned to individual instances of a word in the dataset which is, to some extent, to solve half of the diachronic conceptual change problem. In addition, the results might be different for each dataset, and hence it is a time consuming procedure that must be repeated. However, disregarding differences between datasets might penalize certain datasets, and hence experiments, compared to others, e.g., expecting an invention to appear in a dataset at invention time when in fact there might be a delay of decades.

For both methods there is a large difference between expecting to automatically find the first instance of change or expecting to find the change when it has gained enough momentum to be detectable by context-dependent methods. An example of the differences in momentum but also the differences between datasets can be illustrated with the word *computer*. An earlier common usage of this word was in reference to humans (Grier 2005), but the 'computing device' sense has been on the rise since the electro-mechanical analog computer was invented in the early 20th century and came to play an important role in the second world war, and its incidence has been increasing with the growing importance of digital computers. The frequency of the word *computer* in Google N-grams reaches over 0.0001% in 1934 for the German portion, 1943 for the American English, and 1953 for the British English, meaning that a method evaluated on the latter dataset would be penalized by 20 years compared to one evaluated on a German dataset.[34]

Here we should also mention the sociolinguistic construct APPARENT TIME (Magué 2006) and a similar idea which informs much work in corpus-based lexicography. Apparent time rests on the assumption that crucial aspects of our linguistic repertoire reach a stable state at an early age, say around the age of 20, meaning that e.g., dialect studies can address diachronic development by recording age-stratified speaker samples synchronously, so that the language of a 70-year old is supposed to reflect – in time capsule fashion – current usage about 50 years ago. In a similar way, lexicographers assume that some genres are linguistically more conservative than others, and look for first appearances of new words or new word senses in news text rather than in fiction. Today, the intuition of dialectologists and lexicographers would conspire to single out social media texts as the main harbingers of lexical change (e.g., Fišer & Ljubešić 2018).

[34] The word *Rechner* was and is used in German as a synonym of *Computer*.

6.3 Recommended evaluation procedure for diachronic conceptual change

We recommend the following to be included in any evaluation procedure:

1. *Pre-chosen testset*: Compare the results for target words to words from the same frequency bin, or to the average behavior of all words, to reduce frequency bias, for both positive and negative words.

2. *Grounding in the dataset*: Evaluate backwards referral to the original texts, e.g., by looking at randomly chosen *n*-grams or sentences, where the word under investigation occurs.

3. *Grounding in the outside world*: evaluate with respect to the outside world, e.g., dictionaries and encyclopedias. How well does the result correspond to the expected? The correspondence to the expected is particularly important if claims are made about language in general on the basis of results derived from the corpus.

4. Consider conceptually and/or practically what happens if there is too little evidence in the text (for certain time periods) for a word: can meaning change be found?

5. Consider if the information found is present in the data at hand, or if it stems from the pre-trained models, and therefore possibly relates to a source outside of the dataset under investigation.

6. Can different change types be differentiated in theory? In practice? This question should be answered even if the method is not used for differentiated change types in the study.

7. Can the time of change be found?

8. How does the method scale up to more time points? This relates in particular to those that evaluate change on a few, far apart time points.

9. Always declare and give grounds for evaluation judgments: *Yes, we consider this to be correct because …*, or: *No, we consider this instance to be incorrect because ….*

10. Use proper falsifiability and control conditions.

6.4 Falsifiability and control conditions

Dubossarsky et al. (2017) highlight the importance of falsifiability, by devising a simple "sanity check", creating control conditions where no change of meaning would be expected to occur. They reproduced previous studies which have purported to establish laws of semantic change, two proposed by Hamilton, Leskovec, et al. (2016) and one proposed by themselves (Dubossarsky et al. 2015), finding that in the control conditions, they reported that sense change effects largely disappear or become considerably smaller. They use the Google Books English fiction and sample 10 million 5-grams per year randomly from 1900–1999, each bin spanning a decade. Two control corpora are used, one randomly shuffles the 5-grams from all bins equally. The size of the vocabulary stays the same as in the original corpus, but most semantic change should be equally spread over the corpus, and hence not observable, or observable to a much lesser extent. A second control corpus is created by sampling 10 million 5-grams randomly from 1999, for 30 samples. Since all words are sampled from the same year, there should be no observable semantic change. Word representations are created using word counts, PPMI and SVD reduction of the PPMI matrix, and the three laws are evaluated on both the genuine corpus and the shuffled control corpus. All three laws were verified in the genuine corpus but also found again in the shuffled corpus. The three word representations were used with a cosine similarity measure on the second control corpus, the 30 samples drawn from 1999, and while the changed scores are all lower for the control corpus, they are significantly positive, showing that the proposed change measurements are affected by noise. Using analytic proofs, it is shown that the average cosine distance between a word's vectors from two different samples (using count-based representations) is negatively correlated with the word's frequency.

The linguistic literature provides a wealth of fact and even more discussion about possible driving forces behind both linguistic variation in general and linguistic change, typically accompanied by a large number of empirical linguistic examples. As a minimal methodological requirement, it would behoove authors proposing that a computational method can bring new insight to the study of lexical change in language, to demonstrate in a credible way that other kinds of variation have been taken into account by e.g., the experimental setup, which crucially includes choice of appropriate positive *and negative* data. Especially claims that seem to fly in the face of established truths in the field should be extremely carefully grounded in relevant linguistic scholarship. For instance, Hills & Adelman (2015) report a finding that semantically, the vocabulary of American English has developed in the direction of greater concreteness over the last 200 years, which seems to go against a proposed generalization about semantic change, namely that concrete vocabulary tends to be extended with abstract senses (Urban 2015: 383). A closer scrutiny of the methodology of the study reveals some questionable details. Thus, the list of crowd-sourced concreteness ratings compiled by Brysbaert et al. (2012) used in the study provides only one part of speech and one concreteness score per lemma, e.g. *play* in this dataset is only a verb with a concreteness rating of 3.24 (on a 0–5 scale). In a follow-up study Snefjella et al. (2018) approach the same problem using a considerably more methodologically sophisticated and careful approach, but which still raises some questions. Building on work by Hamilton, Clark, et al. (2016), they compute decadal concreteness scores for the COHA corpus (for the period 1850–2000) based on a small set of seed words assumed to have stayed stable in extreme concreteness and abstractness over the whole investigated time period, and find the same trend of increasing concreteness in the corpus over time. As an anecdotal indication of the accuracy of their approach, they list the top 30 concrete and top 30 abstract (text) words that come out of their computation (e.g., *muddy, knives* vs. *exists, doctrine*) and also report statistical correlations between the computed scores and several sets of human ratings, including those of Brysbaert et al. (2012). However, looking at the scatterplots provided by Snefjella et al. (2018: 6), it is clear that the computed scores inflate concreteness compared to the human ratings, and in particular at the more abstract end of the concreteness range.[35] Further, if we POS tag the results[36] we note that many function words (e.g., determiners and prepositions) come out as highly concrete (e.g., *the* is very close to *muddy* for some of the decades), whereas they cluster consistently at the abstract end in the human ratings. The results reported by Hills & Adelman (2015) and Snefjella et al. (2018) are very interesting to a historical linguist and deserve further study, but their studies should be replicated, with clear control conditions informed by awareness of historical linguistic facts, before any secure conclusions can be drawn.

7 Summary, conclusions and research directions

We summarize below the main observations of our survey.

First of all, we note that the field has grown rapidly in the last few years, resulting in a variety of techniques for lexical semantic change detection, ranging from counting approaches over generative models to neural network based

[35] This does not in itself invalidate their result, of course. If this tendency is consistent over time, we are still seeing a diachronic increase in concreteness of the same magnitude that they report.

[36] Their resulting data are available in their entirety at http://kupermanreadlab.mcmaster.ca/kupermanreadlab/downloads/concreteness-scores.zip

word embeddings. The state of the art is represented by methods based on word embedding techniques. However, most of these approaches are sense-agnostic, effectively focusing on the mixture of word senses expressed by a lexeme. Although some claim that their methods utilize the dominant word sense, they use each occurrence of the lexeme or word form without detecting if it is indeed representing the dominant sense or not.

Another common shortcoming is that only a few approaches propose techniques capable of analyzing semantic change in words with relatively few occurrences. The amount of data for low-frequency words may be insufficient to construct reliable hypotheses using standard methods. Dynamic embeddings seem to offer a more suitable alternative with respect to small datasets. When moving to sense-differentiated embeddings, most likely even more data is needed, and the dynamic embeddings can be a path forward. In relation to this, a common restriction of the discussed methods is that they work on a vocabulary common to all the investigated time periods and make use of the *k* most common words. In some cases, the word frequencies are first normalized per year to avoid a dominance of modern words (since the available digital datasets grow in size over time). Still, this means that only words extant in the datasets over the entire time period contribute to the analysis, both in that they are the only words for which change can be detected, but also because they cannot contribute to the meaning of present words. A word like the Old Swedish legal term *bakvaþi*, meaning ‘the act of accidentally stabbing someone standing behind you when taking aim to swing your sword forward’, is only valid for a period and then disappears from our vocabulary. By ignoring this word, we will not capture any changes regarding the word, which has a very interesting story, but we also prevent it from contributing to the meaning of any of our other *k* words.

In addition, since most of the corpora are not first standardized with respect to spelling variation, many common words are ignored only because their spelling has changed over time. For example, *infynyt, infinit, infinyte, infynit, infineit* are all historical spelling variations used at different times for the word now spelled *infinite* (Simpson & Weiner 1989). To properly address the problem of discovering and describing language change, we need to combine spelling variation, sense variation and lexical replacements in one framework.

Next, while a sense change may be successfully detected *as a diachronic process*, determining the exact time point of semantic change requires the formulation of auxiliary hypotheses about the criteria to be used for determining this. Such criteria are obviously dependent on the available data. For most historical periods we have only texts typically produced by a small and skewed sample of the entire language community. Will thresholds of occurrence in historical texts faithfully reflect underlying change points?

When it comes to evaluating methods and systems, there is a general lack of standardized evaluation practices. Different papers use different datasets and testset words, making it difficult or impossible to compare the proposed solutions. Proper evaluation metrics for semantic change detection and temporal analog detection have not been yet established. Furthermore, comparing methods proposed by different groups is difficult due to varying preprocessing details. For example, filtering out infrequent words can impact the results considerably and different papers employ different thresholds for removing rare words (e.g., some filter out words that appear less than 5 times, others less than 200 times). We suggest the use of SemEval-2020 Task 1 (Schlechtweg et al. 2020) and corresponding, standardized sources and tasks that facilitate comparability, and encourage authors to release their code for better reproducability and model comparison.

Only a few proposals seem to allow for automatically outputting evidence of change to explain to users the nuances of the sense change and to provide concrete examples. Change type determination by automatic means is one step towards this. Related to this is the need for more user-friendly and extensive visualization approaches for diachronic conceptual change analysis given its inherent complexity (see Jatowt et al. 2021). One should keep in mind that many researchers in, for example, the humanities will not accept tools that require programming skills on the part of the user, yet they require tools that are powerful enough to address non-trivial questions and to enable in-depth investigation.

The issue of interdependence between semantic changes of different words is also an interesting avenue of research. Most of the surveyed approaches focus on single words, with only a few authors proposing to view sense change of a target word in relation to another reference word. Future approaches may take entire concepts or topics for investigation so that sense fluctuations of a given word would be seen in the context of changes of other words that may represent the same concept, the same topic or may be semantically related in some other way. Rather than analyzing diachronic conceptual change independently from the changes of other words, a more exhaustive approach could consider also senses of words belonging to an intricate net of word-to-word inter-relations. This could result in a more complete and accurate understanding of why and how a given word changed its sense.

Finally, we note that the linguistic study of semantic change has traditionally been pursued in the context of single languages or language families, and on limited data sets. In particular, nearly all the proposed approaches in the computational literature reviewed here are applied to English data only, due to the dominant position of English in various respects, which is reflected not least in the limited availability of datasets in other languages. Notably, the need for diachronic corpora in other languages than English has also been emphasized in the mentioned survey by Tang (2018). Even if some "laws" of semantic change have been suggested (e.g., Wilkins 1996, Traugott & Dasher 2001), and general classifications of semantic changes into types have been proposed (see Urban 2015), albeit also questioned (see Fortson 2003), the field is still underdeveloped with regard to its empirical basis. For example, it would be necessary to carefully consider whether the underlying corpus is indeed representative of the given language, and does not introduce any bias towards a particular region, gender, social group, and so on, before making any general claims. Approaches that rely on corroborating results using different datasets could be helpful here, especially if informed by a solid knowledge of linguistic methodology and applied to a significant number of genetically, typologically and geographically diverse languages, allowing for both extension and validation of databases such as the *catalogue of semantic shifts* manually compiled by Zalizniak et al. (2012). How applicable the investigated methods will be to other languages is ultimately an empirical matter, but we see no reasons not to be optimistic in this regard.

In view of the above finding, we list below several recommendations:

- When showing and discussing results in a paper, the authors should provide their viewpoint and justification thereof, whether these results are correct or not, and why.

- Always use some sort of control, be it time-stable words or a control dataset, since in isolation, numbers are not sufficient.

- While there have been several methods proposed so far for automatically detecting semantic change, still there are no solutions for automatically generating the "story of the word". Such story-telling would help to concisely explain how the term changed, perhaps giving also a reason for the change (e.g., a new invention). Automatically detecting the type of change could be seen as the first step towards this goal.

Acknowledgments

Thanks go to the Center for Digital Humanities (CDH), University of Gothenburg and to Kyoto University. A first version of this work was written while Nina Tahmasebi was employed at CDH. Adam Jatowt was employed by Kyoto University when the first version was completed. This work has in part been funded by the project *Towards Computational Lexical Semantic Change Detection* supported by the Swedish Research Council (2019–2022; contract 2018-01184), the project *South Asia as a linguistic area?* funded by the Swedish Research Council (2015–2020; contract 2014-00969), and *Nationella språkbanken* (the Swedish National Language Bank), a research infrastructure jointly financed by the Swedish Research Council (2018–2024; contract 2017-00626) and the ten partner institutions. In addition we are grateful for the strong and enduring commitment that the University of Gothenburg, its Faculty of Humanities, and its Department of Swedish have shown to our R&D unit Språkbanken Text for almost 50 years. The authors would also like to acknowledge Dominik Schlechtweg, Haim Dubossarsky, and Simon Hengchen for fruitful discussions and insights. The authors wish to thank the Department of Literature, History of Ideas, and Religion at the University of Gothenburg for providing financial support for language editing services.

Abbreviations

ACL	Association for Computational Linguistics
BNC	British National Corpus
DBSCAN	Density-based spatial clustering of applications with noise
DSG-f	dynamic skip-gram with filtering
DSG-s	dynamic skip-gram with smoothing
EVALITA	Evaluation of NLP and Speech Tools for Italian
LDA	latent Dirichlet allocation
LSA	latent semantic analysis
MRR	mean reciprocal rank
NLP	Natural Language Processing
NMI	normalized mutual information
OCR	optical character recognition
OED	Oxford English Dictionary
PMI	pointwise mutual information
POS	part-of-speech
PPMI	positive pointwise mutual information
SGNS	skip-gram with negative sampling
SVD	singular value decomposition
ukwac	UK web archive
WSI	word sense induction

References

Agirre, Eneko & Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In (SemEval '07), 7–12. ACL. http://dl.acm.org/citation.cfm?id=1621474.1621476.

Aikhenvald, Alexandra & Robert M. W. Dixon. 2002. Word: A typological framework. In Alexandra Aikhenvald & Robert M. W. Dixon (eds.), *Word: A cross-linguistic typology*, 1–41. Cambridge: Cambridge University Press.

Allan, Kathryn & Justyna A. Robinson. 2011. Introduction: Exploring the "state of the art" in historical semantics. In Kathryn Allan & Justyna A. Robinson (eds.), *Current methods in historical semantics*, 1–13. Berlin: De Gruyter Mouton.

Alpher, Barry & David Nash. 1999. Lexical replacement and cognate equilibrium in Australia. *Australian Journal of Linguistics* 19(1). 5–56. DOI: 10.1080/07268609908599573.

Anttila, Raimo. 1972. *An introduction to historical and comparative linguistics.* New York: MacMillan.

Apresjan, Jurij D. 1974. Regular polysemy. *Linguistics* 12(142). 5–32.

Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 1(1). 1–16.

Azarbonyad, Hosein, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx & Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of CIKM 2017*, 1509–1518. Singapore: ACM. DOI: 10.1145/3132847.3132878.

Baker, Adam. 2008. Computational approaches to the study of language change. *Language and Linguistics Compass* 2(2). 289–307.

Bamler, Robert & Stephan Mandt. 2017. Dynamic word embeddings. In Doina Precup & Yee Whye Teh (eds.), *Proceedings of the 34th international conference on machine learning* (Proceedings of Machine Learning Research 70), 380–389. Sydney: PMLR.

Bamman, David & Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries*, 1–10. ACM.

Barkan, Oren. 2017. Bayesian neural word embedding. In *Proceedings of the 31st Conference on Artificial Intelligence*, 3135–3143. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14653.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti & Rossella Varvara. 2020. Overview of the EVALITA 2020 diachronic lexical semantics (DIACR-ita) task. In Valerio Basile, Danilo Croce, Maria Di Maro &

Lucia C. Passaro (eds.), *Proceedings of the 7th evaluation campaign of natural language processing and speech tools for Italian (EVALITA 2020).*

Basile, Pierpaolo, Annalina Caputo, Roberta Luisi & Giovanni Semeraro. 2016. Diachronic analysis of the Italian language exploiting Google Ngram. In *Third Italian Conference on computational Linguistics CLiC-it 2016.*

Basile, Pierpaolo, Annalina Caputo & Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In *First Italian Conference on computational Linguistic CLiC-it 2014.*

Beinborn, Lisa & Rochelle Choenni. 2019. Semantic drift in multilingual representations. *arXiv preprint arXiv:1904.10820.*

Bender, Emily M. 2009. Linguistically Naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, 26–32. Athens: ACL. https://www.aclweb.org/anthology/W09-0106.

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6(3). 1–26.

Bender, Emily M. 2016. Linguistic typology in natural language processing. *Linguistic Typology* 20(3). 645–660.

Berberich, Klaus, Srikanta J. Bedathur, Mauro Sozio & Gerhard Weikum. 2009. *Bridging the terminology gap in Web archive search.* Paper presented at the Twelfth International Workshop on the Web and Databases (WebDB 2009). http://webdb09.cse.buffalo.edu/papers/Paper20/webdb2009-final.pdf.

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

Biemann, Chris. 2006. Chinese whispers: An efficient Graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, 73–80. ACL. http://dl.acm.org/citation.cfm?id=1654758.1654774.

Blei, David M. & John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning*, 113–120.

Borin, Lars. 1988. A computer model of sound change: An example from Old Church Slavic. *Literary and Linguistic Computing* 3(2). 105–108.

Bowern, Claire. 2019. Semantic change and semantic stability: Variation is key. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 48–55.

Boyd-Graber, Jordan, David Blei & Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 EMNLP-CoNLL*, 1024–1033. http://www.aclweb.org/anthology/D/D07/D07-1109.

Brown, Cecil H., Eric W. Holman, Søren Wichmann & Viveka Velupillai. 2008. Automated classification of the world's languages: A description of the method and preliminary results. *Sprachtypologie und Universalienforschung* 61(4). 285–308.

Brysbaert, Marc, Boris New & Emmanual Keuleers. 2012. Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods* 44(4). 991–997.

Bullinaria, John & Joseph Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods* 44. 890–907.

Campbell, Lyle. 2004. *Historical linguistics.* Cambridge: MIT Press.

Cathcart, Chundra A. 2020. A probabilistic assessment of the Indo-Aryan Inner–Outer hypothesis. *Journal of Historical Linguistics* 10(1). 42–86. DOI: 10.1075/jhl.18038.cat.

Ci, Jiwei. 2008/1987. Synonymy and polysemy. In Patrick Hanks (ed.), *Lexicology: Critical concepts in linguistics. Vol. III: Core meaning, extended meaning*, 191–207. Reprinted from Lingua 72 (1987): 315–331. London: Routledge.

Clear, Jeremy. 1992. Corpus sampling. In Gerhard Leitner (ed.), *New directions in English language corpora*, 21–31. Berlin: Mouton de Gruyter.

Cook, Paul, Jey Han Lau, Diana McCarthy & Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014: Technical papers*, 1624–1635. Dublin: ACL. https://www.aclweb.org/anthology/C14-1154.

Cook, Paul, Jey Han Lau, M. Rundell, Diana McCarthy & Timothy Baldwin. 2013. A lexicographic appraisal of an automatic approach for detecting new word-senses. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.), *Proceeedings of eLex 2013*, 49–65. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Cooper, Martin C. 2005. A mathematical model of historical semantics and the grouping of word meanings into concepts. *Computational Linguistics* 32(2). 227–248. DOI: 10.1162/0891201054223995.

Costin-Gabriel, C. & T. E. Rebedea. 2014. Archaisms and neologisms identification in texts. In *2014 RoEduNet Conference*, 1–6. DOI: 10.1109/RoEduNet-RENAM.2014.6955312.

Cruse, D. Alan. 1986. *Lexical semantics.* Cambridge: Cambridge University Press.

Dagan, Ido & Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics* 20(4). 563–596. http://www.aclweb.org/anthology/J94-4003.

Dagan, Ido, Alon Itai & Ulrike Schwall. 1991. Two languages are more informative than One. In *Proceedings of ACL 1991*, 130–137. Berkeley: ACL. http://www.aclweb.org/anthology/P91-1017.

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora* 7(2). 121–157.

Degaetano-Ortlieb, Stefania & Jannik Strötgen. 2017. Diachronic variation of temporal expressions in scientific writing through the lens of relative entropy. In *International Conference of the German Society for Computational Linguistics and Language Technology*, 259–275. Springer.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019: Volume 1 (Long and short papers)*, 4171–4186. Minneapolis: ACL. DOI: 10.18653/v1/N19-1423.

Dorow, Beate, Jean-pierre Eckmann & Danilo Sergi. 2005. Using curvature and markov clustering in graphs for lexical acquisition and word sense discrimination. In *MEANING-2005 Workshop*.

Dridan, Rebecca & Stephan Oepen. 2012. Tokenization: Returning to a long solved problem. A survey, contrastive experiment, recommendations, and toolkit. In *Proceedings of ACL 2012*, 378–382.

Duan, Yijun, Adam Jatowt, Sourav S. Bhowmick & Masatoshi Yoshikawa. 2019. Mapping entity sets in news archives across time. *Data Science and Engineering* 4(3). 208–222.

Dubossarsky, Haim. 2018. *Semantic change at large: A computational approach for semantic change research*. Hebrew University of Jerusalem, Edmond & Lily Safra Center for Brain Sciences. (Doctoral dissertation).

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of ACL 2019*, 457–470. Florence: ACL. DOI: 10.18653/v1/P19-1044.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer & Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of NetWordS 2015* (CEUR Workshop Proceedings 1347), 66–70. Pisa.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Eckart, Carl & Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1. 211–218.

Eger, Steffen & Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of ACL 2016 (Volume 2: Short papers)*, 52–58. Berlin: ACL. DOI: 10.18653/v1/P16-2009.

Eisenstein, Jacob. 2015. Systematic patterning in phonologically-motivated spelling variation. *Journal of Sociolinguistics* 19(2). 161–188.

Eisenstein, Jacob. 2019. Measuring and modeling language change. In *Proceedings of NAACL 2019: Tutorials*, 9–14. ACL.

Ellison, T. Mark & Luisa Miceli. 2017. Language monitoring in bilinguals as a mechanism for rapid lexical divergence. *Language* 93(2). 255–287.

Englhardt, Adrian, Jens Willkomm, Martin Schäler & Klemens Böhm. 2020. Improving semantic change analysis by combining word embeddings and word frequencies. *International Journal on Digital Libraries* 21(3). 247–264.

Erben Johansson, Niklas, Andrey Anikin, Gerd Carling & Arthur Holmer. 2020. The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology* 24(2). 253–310.

Erk, Katrin. 2010. What is word meaning, really? (And how can distributional models help us describe it?) In *GEMS 2010*, 17–26. ACL.

Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Ferraresi, Adriano, Eros Zanchetta, Marco Baroni & Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*.

Ferrari, Alessio, Beatrice Donati & Stefania Gnesi. 2017. Detecting domain-specific ambiguities: An NLP approach based on Wikipedia crawling and word embeddings. In *Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference Workshops*, 393–399.

Firth, John R. 1957. *Papers in linguistics 1934–1951*. London: Oxford University Press.

Fišer, Darja & Nikola Ljubešić. 2018. Distributional modelling for semantic shift detection. *International Journal of Lexicography* 32(2). 1–21. DOI: 10.1093/ijl/ecy011.

Fortson, Benjamin W., IV. 2003. An approach to semantic change. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 648–666. Oxford: Blackwell.

François, Alexandre. 2008. Semantic maps and the typology of colexication: Intertwining polysemy networks across languages. In Martine Vanhove (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 163–215. Amsterdam: John Benjamins.

Frermann, Lea & Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the ACL* 4. 31–45. DOI: 10.1162/tacl_a_00081.

Gale, William A., Kenneth W. Church & David Yarowsky. 1992. One sense per discourse. In *Speech and natural language: Proceedings of a workshop held at Harriman, New York, February 23–26, 1992*. https://www.aclweb.org/anthology/H92-1045.

Giulianelli, Mario, Marco Del Tredici & Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of ACL 2020*, 3960–3973. Online: ACL. DOI: 10.18653/v1/2020.acl-main.365.

Grier, David Alan. 2005. *When computers were human*. Princeton: Princeton University Press.

Gulordava, Kristina & Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 67–71. Edinburgh: ACL. https://www.aclweb.org/anthology/W11-2508.

Hamilton, William L., Kevin Clark, Jure Leskovec & Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of EMNLP 2016*, 595–605. ACL.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Hanks, Patrick. 2013. *Lexical analysis: Norms and exploitations*. Cambridge: MIT Press.

Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica* 45(1). 31–80.

Heine, Bernd & Tania Kuteva. 2002. *World lexicon of grammaticalization*. Cambridge: Cambridge University Press.

Henrich, Joseph, Steven J. Heine & Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33. 61–135.

Hewson, John. 1973. Reconstructing prehistoric languages on the computer: The triumph of the electronic neogrammarian. In *Proceedings of the 5th conference on computational linguistics*, 263–273. Pisa: ACL.

Hewson, John. 1974. Comparative reconstruction on the computer. In John M. Anderson & Charles Jones (eds.), *Historical Linguistics: Proceedings of the First International Conference on Historical Linguistics*, 191–197. Amsterdam: North-Holland.

Hill, Felix, Roi Reichart & Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* 41(4). 665–695.

Hills, Thomas T. & James S. Adelman. 2015. Recent evolution of learnability in American English from 1800 to 2000. *Cognition* 143. 87–92.

Holzmann, Helge, Gerhard Gossen & Nina Tahmasebi. 2012. Fokas: Formerly known as, a search engine incorporating named entity evolution. In *Proceedings of COLING 2012: Demonstration papers*, 215–222.

Holzmann, Helge & Thomas Risse. 2014. Insights into entity name evolution on Wikipedia. In *Proceedings of WISE 2014*, vol. 8787, 47–61. Springer. DOI: 10.1007/978-3-319-11746-1_4.

Holzmann, Helge, Nina Tahmasebi & Thomas Risse. 2015. Named entity evolution recognition on the Blogosphere. *International Journal on Digital Libraries* 15(2–4). 209–235.

Hopper, Paul J. & Elizabeth Closs Traugott. 1993. *Grammaticalization.* Cambridge: Cambridge University Press.

Hu, Renfen, Shen Li & Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of ACL 2019*, 3899–3908. Florence: ACL. DOI: 10.18653/v1/P19-1379.

Ihalainen, Pasi. 2006. Between historical semantics and pragmatics. *Journal of Historical Pragmatics* 7(1). 115–143.

Jatowt, Adam, Ricardo Campos, Sourav S. Bhowmick, Nina Tahmasebi & Antoine Doucet. 2018. Every word has its history: Interactive exploration and Visualization of word sense evolution. In *Proceedings of CIKM 2018*, 1899–1902. ACM.

Jatowt, Adam, Nina Tahmasebi & Lars Borin. 2021. Computational approaches to lexical semantic change: Visualization systems and novel applications. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 311–339. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040320.

Johansson, Stig. 1994. ICAME – Quo vadis? Reflections on the use of computer corpora in linguistics. *Computers and the Humanities* 28(4–5). 243–252.

Johnson, Mark. 1985. Computer aids for comparative dictionaries. *Linguistics* 23(2). 285–302.

Joseph, Brian D. & Richard D. Janda (eds.). 2003. *The handbook of historical linguistics.* Oxford: Blackwell.

Kahmann, Christian, Andreas Niekler & Gerhard Heyer. 2017. Detecting and assessing contextual change in diachronic text documents using context volatility. In *Proceedings of the 9th International Joint Conference on Knowledge Discovery*, 135–143. DOI: 10.5220/0006574001350143.

Kaluarachchi, Amal C., Debjani Roychoudhury, Aparna S. Varde & Gerhard Weikum. 2011. SITAC: Discovering semantically identical temporally altering concepts in text archives. In *Proceedings of EDBT/ICDT 2011*, 566–569. DOI: 10.1145/1951365.1951442.

Kaluarachchi, Amal C., Aparna S. Varde, Srikanta Bedathur, Gerhard Weikum, Jing Peng & Anna Feldman. 2010. Incorporating terminology evolution for query translation in text retrieval with association rules. In *Proceedings of the 19th ACM international conference on information and knowledge management*, 1789–1792.

Kanhabua, Nattiya & Kjetil Nørvåg. 2010. Exploiting time-based synonyms in searching document archives. In *JCDL*, 79–88. DOI: 10.1145/1816123.1816135.

Karjus, Andres, Richard A. Blythe, Simon Kirby & Kenny Smith. 2020. *Communicative need modulates competition in language change.* https://arxiv.org/abs/2006.09277.

Kenter, Tom, Melvin Wevers, Pim Huijnen & Maarten De Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1191–1200.

Kerremans, Daphné, Susanne Stegmayr & Hans-Jörg Schmid. 2011. The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In Kathryn Allan & Justyna A. Robinson (eds.), *Current methods in historical semantics*, 59–96. Berlin: De Gruyter Mouton.

Kilgarriff, Adam. 1997. "I don't believe in word senses". *Computers and the Humanities* 31(2). 91–113.

Kilgarriff, Adam. 2004. How common is the commonest sense of a word? In Ivan Kopeček & Karel Pala (eds.), *Proceedings of TSD 2004*, 1–9. Berlin: Springer.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore: ACL. DOI: 10.3115/v1/W14-2517.

Koplenig, Alexander. 2016. *Analyzing lexical change in diachronic corpora*. Universität Mannheim. (Doctoral dissertation). http://nbn-resolving.de/urn:nbn:de:bsz:mh39-48905.

Koplenig, Alexander. 2017a. A data-driven method to identify (correlated) changes in chronological corpora. *Journal of Quantitative Linguistics* 24(4). 289–318.

Koplenig, Alexander. 2017b. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data set – Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities* 32(1). 169–188.

Koplenig, Alexander & Carolin Müller-Spitzer. 2016. Population size predicts lexical diversity, but so does the mean sea level–why it is important to correctly account for the structure of temporal data. *PLOS ONE* 11(3). e0150771.

Koptjevskaja-Tamm, Maria. 2008. Approaching lexical typology. In Martine Vanhove (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 3–52. Amsterdam: John Benjamins.

Koptjevskaja-Tamm, Maria. 2012. New directions in lexical typology. *Linguistics* 50(3). 373–394.

Koptjevskaja-Tamm, Maria, Ekaterina Rakhilina & Martine Vanhove. 2016. The semantics of lexical typology. In Nick Riemer (ed.), *The Routledge handbook of semantics*, 434–554. London: Routledge.

Kroch, Anthony S. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1(3). 199–244. DOI: 10.1017/S0954394500000168.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey. 2020. *Distributional word embeddings in modeling diachronic semantic change*. University of Oslo. (Doctoral dissertation).

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*, 1384–1397. Santa Fe: ACL.

Lange, Sven. 2002. Betydelsestrukturen hos polysema ord i SAOB [The semantic structure of polysemous words in the Swedish Academy dictionary]. In *Strövtåg i nordisk språkvetenskap*, 7–20. Berlin: Nordeuropa-Institut der Humboldt-Universität zu Berlin.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman & Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of EACL 2012*, 591–601. Avignon: ACL. https://www.aclweb.org/anthology/E12-1060.

Levy, Omer, Yoav Goldberg & Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL* 3. 211–225.

Liberman, Mark. 2013. *~~Word~~ String frequency distributions*. Blog posting on Language Log, 2013-02-03. http://languagelog.ldc.upenn.edu/nll/?p=4456.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449(7163). 713.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL 1998*, 768–774. ACL. DOI: 10.3115/980691.980696.

Linell, Per. 2005/1982. *The written language bias in linguistics: Its nature, origins and transformations*. First published in 1982 by Linköping University, Dept. of Communication Studies. London: Routledge.

Lowe, John B. & Martine Mazaudon. 1994. The reconstruction engine: A computer implementation of the comparative method. *Computational Linguistics* 20(3). 381–417.

Magué, Jean-Philippe. 2006. Semantic changes in apparent time. In *Proceedings of the thirty-second annual meeting of the Berkeley Linguistics Society*, 227–235. Berkeley Linguistics Society.

Mailhammer, Robert. 2015. Etymology. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 423–441. London: Routledge.

Malkiel, Yakov. 1993. *Etymology*. Cambridge: Cambridge University Press.

Martinc, Matej, Petra Kralj Novak & Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of LREC 2020*, 4811–4819. Marseille: ELRA. https://www.aclweb.org/anthology/2020.lrec-1.592.

Martinc, Matej, Syrielle Montariol, Elaine Zosa & Lidia Pivovarova. 2020. Capturing evolution in word usage: Just add more clusters? In Amal El Fallah Seghrouchni, Gita Sukthankar, Tie-Yan Liu & Maarten van Steen (eds.), *Companion of the 2020 Web Conference*, 343–349. Taipei: ACM. DOI: 10.1145/3366424.3382186.

Matthews, Peter H. 1974. *Morphology: An introduction to the theory of word-structure*. Cambridge: Cambridge University Press.

Mei, Qiaozhu & ChengXiang Zhai. 2005. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In (KDD '05), 198–207. ACM. DOI: 10.1145/1081870.1081895.

Meringer, Rudolf. 1912. Zur Aufgabe und zum Namen unserer Zeitschrift [about the mission and the name of our magazine]. *Wörter und Sachen: Kulturhistorische Zeitschrift für Sprach- und Sachforschung* 3. 22–56.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331. 176–182.

Mihalcea, Rada & Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of ACL 2012: Short papers - Vol. 2*, 259–263. ACL. http://dl.acm.org/citation.cfm?id=2390665.2390727.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint 1301.3781*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in neural information processing systems*, 3111–3119. Red Hook, NY: Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal & Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21(5). 773–798.

Mitra, Sunny, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee & Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of ACL 2014 (Volume 1: Long papers)*, 1020–1029. Baltimore: ACL. DOI: 10.3115/v1/P14-1096.

Montariol, Syrielle & Alexandre Allauzen. 2019. Empirical study of diachronic word embeddings for scarce data. In Ruslan Mitkov & Galia Angelova (eds.), *Proceedings of RANLP 2019*, 795–803. Varna: INCOMA Ltd. DOI: 10.26615/978-954-452-056-4_092.

Morsy, Sara & George Karypis. 2016. Accounting for language changes over time in document similarity search. *ACM Transactions on Information Systems* 35(1).

Murphy, Gregory L. 2002. *The big Book of concepts*. Cambridge: MIT Press.

Murphy, M. Lynne. 2003. *Semantic relations and the lexicon*. Cambridge: Cambridge University Press.

Navigli, Roberto. 2012. A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012*, 115–129. Springer.

Newman, John. 2016. Semantic shift. In Nick Riemer (ed.), *The Routledge handbook of semantics*, 266–280. London: Routledge.

Ohshima, Hiroaki & Katsumi Tanaka. 2010. High-speed detection of ontological knowledge and bi-directional lexico-syntactic patterns from the web. *JSW* 5(2). 195–205.

Orlikowski, Matthias, Matthias Hartung & Philipp Cimiano. 2018. Learning diachronic analogies to analyze concept change. In *Proceedings of the second joint sighum workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, 1–11. Santa Fe: ACL. https://www.aclweb.org/anthology/W18-4501.

Östling, Robert. 2016. Studying colexification through massively parallel corpora. In Päivi Juvonen & Maria Koptjevskaja-Tamm (eds.), *The lexical typology of semantic shifts*, 157–176. Amsterdam: De Gruyter Mouton.

Pagel, Mark, Quentin D. Atkinson & Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163). 717.

Pechenick, Eitan Adam, Christopher M. Danforth & Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE* 10(10). e0137041.

Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith & Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 56–66. Florence: ACL. DOI: 10.18653/v1/W19-4707.

Petersen, Alexander M., Joel N. Tenenbaum, Shlomo Havlin, H. Eugene Stanley & Matjaž Perc. 2012. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports* 2(943). 1–10. DOI: 10.1038/srep00943.

Ramiro, Christian, Mahesh Srinivasan, Barbara C. Malt & Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences* 115. 2323–2328.

Recchia, Gabriel, Ewan Jones, Paul Nulty, John Regan & Peter de Bolla. 2016. Tracing shifting conceptual vocabularies through time. In *European knowledge acquisition workshop*, 19–28. Springer.

Richter, Melvin. 1996. Appreciating a contemporary classic: The Geschichtliche Grundbegriffe and future scholarship. In Hartmut Lehmann & Melvin Richter (eds.), *The meaning of historical terms and concepts: New studies on Begriffsgeschichte*, 7–20. Washington: German Historical Institute.

Riemer, Nick. 2010. *Introducing semantics.* Cambridge: Cambridge University Press.

Rodda, Martina A., Marco S.G. Senaldi & Alessandro Lenci. 2016. Panta rei: Tracking semantic change with distributional semantics in ancient Greek. In Pierpaolo Basile et al. (eds.), *Proceedings of third Italian conference on Computational Linguistics (CLiC-it 2016)* (CEUR Workshop Proceedings 1749). http://ceur-ws.org/Vol-1749/paper46.pdf.

Rodda, Martina A., Marco S.G. Senaldi & Alessandro Lenci. 2017. Panta rei: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics* 3(1). 11–24. https://arpi.unipi.it/handle/11568/891899.

Rodina, Julia, Daria Bakshandaeva, Vadim Fomin, Andrey Kutuzov, Samia Touileb & Erik Velldal. 2019. Measuring diachronic evolution of evaluative adjectives with word embeddings: The case for English, Norwegian, and Russian.

In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 202–209.

Roget, Mark Peter. 1852. *Thesaurus of English words and phrases*. London: Longman.

Rosenfeld, Alex & Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of NAACL-HLT 2018: Volume 1 (Long papers)*, 474–484. New Orleans: ACL. DOI: 10.18653/v1/N18-1044.

Rudolph, Maja R. & David M. Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of WWW 2018*, 1003–1011. ACM. DOI: 10.1145/3178876.3185999.

Ryskina, Maria, Ella Rabinovich, Taylor Berg-Kirkpatrick, David R. Mortensen & Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. *Proceedings of the Society for Computation in Linguistics* 3(1). 43–52.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 104–111. Athens: ACL. https://www.aclweb.org/anthology/W09-0214.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2011. Tracing semantic change with Latent Semantic Analysis. In Kathryn Allan & Justyna A. Robinson (eds.), *Current methods in historical semantics*, 162–183. Berlin: De Gruyter Mouton.

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici & Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of ACL 2019*, 732–746. Florence: ACL.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of SemEval 2020*, 1–23. Barcelona: ACL. https://www.aclweb.org/anthology/2020.semeval-1.1.

Schlechtweg, Dominik & Sabine Schulte im Walde. 2020. Simulating lexical semantic change from sense-annotated data. In C. Cuskley, M. Flaherty, H. Little, Luke McCrohon, A. Ravignani & T. Verhoef (eds.), *The Evolution of Language: Proceedings of the 13th International Conference (EVOLANGXIII)*.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–123.

Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale & Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of EMNLP-IJCNLP 2019*, 66–76. Hong Kong: ACL.

Simpson, John & Edmund Weiner (eds.). 1989. *The Oxford English Dictionary*. 2nd edn. http://dictionary.oed.com.

Smith, K. Aaron. 2011. Grammaticalization. *Language and Linguistics Compass* 5(6). 367–380.

Snefjella, Bryor, Michel Généreaux & Victor Kuperman. 2018. Historical evolution of concrete and abstract language revisited. *Behavior Research Methods* online first. 1–13. DOI: 10.3758/s13428-018-1071-2.

Søgaard, Anders. 2016. Biases we live by. *Nordisk Tidsskrift for Informationsvidenskab og Kulturformidling* 5(1). 31–35.

Stewart, Ian, Dustin Arendt, Eric Bell & Svitlana Volkova. 2017. Measuring, predicting and visualizing short-term change in word representation and usage in VKontakte social network. In *Proceedings of ICWSM 2017*, 672–675. Montréal: AAAI Press.

Szymanski, Terrence. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of ACL 2017*. ACL.

Taavitsainen, Irma & Susan Fitzmaurice. 2007. Historical pragmatics: What it is and how to do it. In Susan Fitzmaurice & Irma Taavitsainen (eds.), *Methods in historical pragmatics*, 11–36. Berlin: Mouton de Gruyter.

Tahmasebi, Nina. 2013. *Models and algorithms for automatic detection of language evolution*. Gottfried Wilhelm Leibniz Universität Hannover. (Doctoral dissertation). http://edok01.tib.uni-hannover.de/edoks/e01dh13/771705034.pdf.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278*.

Tahmasebi, Nina, Lars Borin, Adam Jatowt & Yang Xu (eds.). 2019. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Florence: ACL. https://www.aclweb.org/anthology/W19-4700.

Tahmasebi, Nina, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann & Thomas Risse. 2012. NEER: An unsupervised method for named entity evolution recognition. In *Proceedings of COLING 2012*, 2553–2568. Mumbai: ACL. https://www.aclweb.org/anthology/C12-1156.

Tahmasebi, Nina, Tereza Iofciu, Thomas Risse, Claudia Niederee & Wolf Siberski. 2008. Terminology evolution in web archiving: Open issues. In *8th International Web Archiving Workshop in conjunction with ECDL 2008, Aarhus, Denmark*.

Tahmasebi, Nina, Kai Niklas, Gideon Zenz & Thomas Risse. 2013. On the applicability of word sense discrimination on 201 years of modern English. *International Journal on Digital Libraries* 13. 135–153. DOI: 10.1007/s00799-013-0105-8.

Tahmasebi, Nina & Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of RANLP 2017*, 741–749. Varna: INCOMA Ltd. DOI: 10.26615/978-954-452-049-6_095.

Takamura, Hiroya, Ryo Nagata & Yoshifumi Kawasaki. 2017. Analyzing semantic change in Japanese loanwords. In *EACL*, 1195–1204. ACL. http://aclweb.org/anthology/E17-1112.

Tang, Xuri. 2018. Survey paper: A state-of-the-art of semantic change computation. *Natural Language Engineering* 24(5). 649–676.

Tang, Xuri, Weiguang Qu & Xiaohe Chen. 2013. Semantic change computation: A successive approach. In *Behavior and social computing*, 68–81. Springer.

Tang, Xuri, Weiguang Qu & Xiaohe Chen. 2016. Semantic change computation: A successive approach. *World Wide Web* 19(3). 375–415. DOI: 10.1007/s11280-014-0316-y.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal & David M. Blei. 2004. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS 2004*, 1385–1392. http://papers.nips.cc/paper/2698-sharing-clusters-among-related-groups-hierarchical-dirichlet-processes.

Tjong Kim Sang, Erik. 2016. Finding rising and falling words. In *Proceedings of LT4DH 2016*, 2–9. Osaka: ACL. http://aclweb.org/anthology/W16-4002.

Traugott, Elizabeth Closs. 2017. Semantic change. In *Oxford research encyclopedia of linguistics*. Oxford: Oxford University Press. (26 June, 2018).

Traugott, Elizabeth Closs & Richard B. Dasher. 2001. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Turney, Peter D. & Saif M. Mohammad. 2019. The natural selection of words: Finding the features of fitness. *PLOS ONE* 14(1). e0211512.

Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1). 141–188.

Ullmann, Stephen. 1959. *The principles of semantics*. London: Blackwell.

Underwood, Ted (ed.). 2019. *Distant horizons: Digital evidence and literary change*. Chicago: University of Chicago Press.

Urban, Matthias. 2015. Lexical semantic change and semantic reconstruction. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 374–392. London: Routledge.

van Aggelen, Astrid, Laura Hollink & Jacco van Ossenbruggen. 2016. Combining distributional semantics and structured data to study lexical change. In *Drift-a-LOD*.

Vejdemo, Susanne & Thomas Hörberg. 2016. Semantic factors predict the rate of lexical replacement of content words. *PLOS ONE* 11(1). e0147924.

Viola, Lorella & Jaap Verheul. 2020. One Hundred Years of migration discourse in the Times: A discourse-historical word vector space approach to the construction of meaning. *Frontiers in Artificial Intelligence* 3. 64. DOI: 10.3389/frai.2020.00064.

Wang, Jing, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart & T. Yu Clement. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. *TACL* 3. 59–71.

Wang, Xuerui & Andrew McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, 424–433. ACM. DOI: 10.1145/1150402.1150450.

Wijaya, Derry Tanti & Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of DETECT 2011*, 35–40. ACM.

Wilkins, David P. 1996. Natural tendencies of semantic change and the search for cognates. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed: Regularity and irregularity in language change*, 264–304. Oxford: OUP.

Wilks, Yorick. 2009. Ontotherapy, or how to stop worrying about what there is. In Nicolas Nicolov, Galia Angelova & Ruslan Mitkov (eds.), *RANLP 2009*, 1–20. Borovets, Bulgaria: Association for Computational Linguistics.

Wishart, Ryder A. 2018. Hierarchical and distributional lexical field theory: A critical and empirical development of Louw and Nida's semantic domain model. *International Journal of Lexicography* 2018/advance articles(ecy015). 1–26. DOI: 10.1093/ijl/ecy015.

Xu, Yang & Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual meeting of the Cognitive Science Society, CogSci 2015*. Pasadena.

Xu, Yang, Barbara C. Malt & Mahesh Srinivasan. 2017. Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive psychology* 96. 41–53.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao & Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 673–681.

Zalizniak, Anna A., Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak & Maxim Russo. 2012. The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics* 50(3). 633–669.

Zhang, Yating, Adam Jatowt, S. Sourav Bhowmick & Yuji Matsumoto. 2019. ATAR: Aspect-based temporal analog retrieval system for document archives. In *WSDM2019*. Melbourne: ACM.

Zhang, Yating, Adam Jatowt, Sourav S. Bhowmick & Katsumi Tanaka. 2016. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering* 28(10). 2793–2807.

Zhang, Yating, Adam Jatowt & Katsumi Tanaka. 2016. Towards understanding word embeddings: Automatically explaining similarity of terms. In *International Conference on Big Data 2016*, 823–832. IEEE.

Zhang, Yating, Adam Jatowt & Katsumi Tanaka. 2017. Temporal analog retrieval using transformation over dual hierarchical structures. In (CIKM '17), 717–726. Singapore: ACM. DOI: 10.1145/3132847.3132917.

Zhou, Jinan & Jiaxin Li. 2020. TemporalTeller at SemEval-2020 Task 1: Unsupervised lexical semantic change detection with temporal reference. In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Barcelona: ACL.

Chapter 2

Semantic changes in harm-related concepts in English

Ekaterina Vylomova & Nick Haslam
The University of Melbourne

The chapter investigates semantic changes in core concepts of psychology, specifically focusing on those related to harm. Haslam (2016) hypothesized that many psychological concepts associated with harm (i.e., forms of psychological disturbance, threat, and maltreatment) have undergone semantic broadening in the past half-century in association with cultural shifts and social change. The implications of this "concept creep" hypothesis have been previously explored by prominent social, political, and legal thinkers (Levari et al. 2018, Lukianoff & Haidt 2019, Pinker 2018, Sunstein 2018), but its linguistic dimension has received little empirical attention.

Here we apply computational models in order to address the concept creep hypothesis. We start with a description of a typology of semantic shifts and provide a summary of computational methods for automatic detection of the most common changes (broadening, narrowing, hyperbole, and litotes) and utilise those to evaluate core harm-related concepts such as 'trauma', 'harassment', and 'bullying' on a new corpus of psychology literature extending from 1970 to 2017. Our results confirm the initial hypothesis and are in line with earlier studies: most concepts became broader and milder over the last few decades. We then continue with a more detailed study in order to understand how exactly the concepts changed, and to do so employ and evaluate different types of semantic representations.

Finally, we additionally train the models on a general domain corpus in order to investigate whether the broadening of harm-related concepts also applies to society at large, rather than only to the academic discourse of psychology. Haslam's influential account of concept creep (Haslam 2016) proposes that broadened concepts of harm disseminate from academic language into wider public use. This final analysis enables a direct test of that conjecture, including comparative analysis of the extent and timing of historical semantic changes across the two corpora.

Ekaterina Vylomova & Nick Haslam. 2021. Semantic changes in harm-related concepts in English. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 93–121. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040304

1 Introduction

Recent years witnessed significant progress in many downstream tasks in natural language processing (NLP) such as machine translation, part-of-speech tagging, language modelling, and many others.[1] Unlike earlier machine learning models that were often provided with a set of pre-designed features or rules, most recent models inherently "learn" them from raw data in the form of dense vectors (embeddings). Training strategies used in the models to learn the embeddings often rely on the distributional semantics hypothesis that states that a word's meaning can be expressed as a distribution over a set of its contexts (Firth 1957, Harris 1954, Weaver 1955). A significant amount of research works explored what aspects of language are captured in these representations. Although the distributional semantics approach presents certain limitations (Bender & Koller 2020), it still allows to extract a surprising amount of information about semantic, morphological, and syntactic properties of language (Mikolov, Yih, et al. 2013, Vylomova et al. 2016, Gladkova et al. 2016, Belinkov & Glass 2019, Rogers et al. 2020). In addition, representations obtained using this approach capture associations between words and can potentially simulate surveys on free word associations (Agirre et al. 2009, Antoniak & Mimno 2018). These successes induced a novel direction of interdisciplinary studies – corpus-centered research – where embeddings are used as a direct evidence about the language and culture of the authors of a training corpus (Antoniak & Mimno 2018). For instance, Hamilton et al. (2016a,b) presented one of the earliest diachronic language models and metrics to evaluate semantic shifts as well as computational approaches to lexical semantic change detection. Over the last few years, the area has significantly increased and witnessed substantial progress and development (Schlechtweg et al. 2020).

In this chapter, we apply diachronic language modelling to computationally attest semantic shifts in core concepts of social psychology. In particular, we focus on diachronic change in the meaning of harm-related concepts and test a "concept creep" hypothesis proposed in Haslam (2016). The hypothesis states that during the past half-century many concepts associated with harm have broadened their meanings in Western societies. We quantitatively evaluate changes in the five negative concepts: 'addiction', 'bullying', 'harassment', 'prejudice', and 'trauma'. We attest them on a newly introduced corpus of psychology journal abstracts and a general domain corpus comprising CoCA and CoHA. In order to test the hypothesis, we first conduct frequency-based analysis and then study the changes in a greater detail by evaluating vector representations learned by epoch-specific models trained on each corpus.

[1] See https://nlpprogress.com/ for most recent state-of-the-art models in each task.

2 The notion of concept creep

Haslam (2016) introduced the idea of "concept creep" to describe a general pattern of semantic inflation in several fundamental psychological concepts. The paper presented a series of case studies in which psychological researchers and theorists expanded the sense of harm-related concepts by loosening definitions to include milder instances ("vertical creep") or by extending definitions to encompass qualitatively new phenomena ("horizontal creep").

The two forms of creep can be understood from the perspective of Bloomfield's typology of lexical semantic change (Bloomfield 1933). Out of seven types identified in the book, some of them are particularly relevant to the current creep study. First, changes may happen along the semantic narrowing (the Old English *mete* 'food' > *meat* 'edible flesh') – widening (the Middle English *briddle* 'young birdling' > *bird* 'birds of all ages') axis. Alternatively, a word's meaning may extend by means of analogy (the Old English *bītan* 'to bite' > the Middle English *bitter* 'acrid').

Indeed, modern studies of word semantics change are based on a long tradition. Yet in the end of the 19th century Bréal (1897) analyzed different types of word meaning change in a diachronical perspective for multiple languages. Particularly, the four major types of concept creep discussed in the current chapter (two vertical and two horizontal ones) were reflected in some form within the taxonomy proposed by Bréal. The horizontal concept broadening is similar to what he referred to as "élargissement de sens" (sense enlargement). One of the examples mentions Latin *pecunia* the meaning of which has gradually broadened from 'richness in possession of livestock' to a general sense of 'wealth'. The vertical broadening deems falling into the "épaississement de sens" category ("sense thickening"). We can notice that in the latter case Bréal mostly speaks about facts of meaning change accompanied by either morphological or non-morphological modification of a word in hand. Thus, a word was not required to keep its exact form, in contrast to the approach we follow in the current study. The phenomenon of concept narrowing was not directly outlined in the Bréal's taxonomy. However, both its horizontal and vertical types seem to be covered by different kinds of metaphor.

Horizontal creep comprises both of these types: widening of 'abuse' to include passive neglect and metaphoric extension of (physical) 'bullying' to include 'cyber-bullying'. Another type of shift might occur along the litotes–hyperbole axis. Litotes represents the change from a weaker to a stronger meaning (the Proto-West Germanic **kwalljan* 'to make suffer' > the Old English *cwellan* 'to kill'), whereas hyperbole is the shift in the opposite direction (the Vulgar Latin

**extonare* 'to strike with thunder' > *astonish* 'to surprise greatly'). This type of change seems to be more pertinent to vertical creep: as we will further show, 'trauma' has transformed to refer to relatively mild adversities (Haslam & McGrath 2020). Horizontal and vertical creep are not mutually exclusive – a concept may change in both ways simultaneously. For example, the concept of 'mental disorder' has progressively broadened in recent decades by relaxing the diagnostic criteria of some conditions (vertical creep; Fabiano & Haslam 2020) and by expanding the range of problems conceptualized as falling within the psychiatric domain (horizontal creep).

Haslam (2016) and Haslam et al. (2020) documented how similar semantic inflation had occurred for the following putatively creeping concepts which we will further examine in the current chapter:

Addiction: This concept originally referred to physiological dependency on an ingested substance, but is increasingly used to identify psychological compulsions to engage in non-ingestive behaviors such as gambling or shopping.

Bullying: This concept, introduced to psychology in the 1970s, initially described peer aggression between children that was repeated, intentional, and perpetrated in the context of a power imbalance. More recent definitions extend bullying to adult workplace settings and relax the repetition, intentionality, and power imbalance criteria.

Harassment: Early uses of this concept emphasized inappropriate sexual approaches but more recently harassment is also used within psychology to refer to nonsexual forms of unwanted attention.

Prejudice: The original psychological definitions of prejudice restricted it to overt animosity towards ethnic or racial outgroups. More recent theory and research extend it to many non-racial groups, allow for covert or nonconscious prejudice, and indicate that it may be manifest as anxiety or condescension rather than hostility. Recent studies showed that it expanded to include subtle micro-aggressions (Lilienfeld 2017).

Trauma: Four decades ago only personally encountered life-threatening events that are outside the realm of normal experience were recognized as traumatic by psychologists and psychiatrists. More recent definitions include vicarious or indirect experiences of stressful events, including those that are relatively prevalent.

Haslam (2016) proposed that these diverse concepts shared a focus on harm (i.e., the experience or infliction of actual or potential suffering). It was further speculated that the correlated broadening of the creeping concepts reflected a rising sensitivity to harm within Western cultures.

3 Related work

We will provide related research for three aspects of our study: the central hypothesis of "concept creep", computational approaches to semantic change detection, and factors that might influence semantic change.

3.1 "Concept creep"

Existing work on concept creep with a few notable exceptions is primarily theoretical and the idea has been taken up by influential writers. Lukianoff & Haidt (2019) have employed it to understand political conflict on college campuses. Pinker (2018) has argued that concept creep leads people to under-estimate social progress because their definitions of hardship expand to include increasingly minor problems. This phenomenon has been demonstrated by Levari et al. (2018), who showed that concept definitions broaden as concept instances become scarcer. McGrath et al. (2019) have explored the attributes of people who hold relatively broad harm-related concepts, finding that they tend to be politically liberal and empathetic, and their personal morality is tied to harm and care for others. Wheeler et al. (2019) studied the Google Books English language corpus and showed that words representing harm-based morality has become more culturally salient (i.e., relatively frequent) in the past four decades, consistent with the theory of concept creep. Most recently, Vylomova et al. (2019) trained a count-based model from Sagi et al. (2009) and a prediction-based one introduced in Hamilton et al. (2016b) on a massive corpus of abstracts of academic psychology journals to evaluate semantic breadth changes in some of the creeping concepts described in Haslam (2016).

3.2 Computational approaches to semantic change detection

Although diachronic studies of language have a long history in linguistics, computational approaches were introduced only recently. Jurgens & Stevens (2009), one of the first, proposed an algorithm for tracking temporal semantic changes by learning a sequence of distributional models over time. The work was followed by an LSA-based model from Sagi et al. (2009). Kim et al. (2014) and Hamilton et al. (2016b) then proposed the first prediction-based neural language models. The training strategies of the models differed, though: Kim et al. (2014) incrementally trained models on each subsequent epoch, while Hamilton et al. (2016b) trained several epoch-specific models independently and then aligned them using Procrustes. Kulkarni et al. (2015) also followed the same direction but only aligned the nearest neighbors rather than the whole space. Both Kulkarni et al. (2015) and Hamilton et al. (2016b) further demonstrated that such prediction-based models (word2vec, in particular) outperform count-based ones on the semantic shifts detection tasks. Further, Dubossarsky et al. (2019) demonstrated that alignment-based diachronic models often introduce additional noise to the representations and proposed a temporal referencing approach that does not require vector space alignment.

3.3 Factors that influence semantic changes

Hamilton et al.'s work in 2016 was influential because they also attempted to state laws of semantic change that would explain the variability in word change rates and identify factors that influence said rates. On the other hand, this research direction was not entirely novel for the scientific community outside of NLP: historical linguistics presents a vast line of work on this topic. For instance, Stern (1931) and Lehrer (1985) suggested that words with close meanings that are strongly associated with one another undergo similar changes ("the law of parallel change"). Contrary to that, Sturtevant (1917) stated "the law of differentiation", i.e. that words with similar meanings (synonyms) tend to diverge over time. Xu & Kemp (2015) evaluated the two laws and provided more evidence for support of "the law of parallel change". Geeraerts et al. (1999) suggested that prototypicality also plays a role: more salient, prototypical meanings will be less likely to change. "The law of prototypicality" was then examined in Dubossarsky et al. (2015), the work demonstrating that the closer a word is to the centroid of the corresponding semantic category cluster, the less likely its meaning changes. Another linguistic hypothesis states that "words become semantically extended by being used in diverse contexts" (Winter et al. 2014) and meaning evolves in a directional fashion: words that have more word associations and senses are more likely to acquire new meanings. Finally, Hamilton et al. (2016b) proposed a hypothesis stating that frequency and polysemy explain most variance in the rates of lexical semantic change. Their study resulted in a more comprehensive understanding of the earlier observations, and resulted in the following two laws of semantic change: (1) "The law of conformity": frequently used words change at slower rates; and (2) "The law of innovation": polysemous words change at faster rates. Later, Dubossarsky et al. (2017) re-considered the laws of semantic change and showed that (1) “the law of innovation” is to a large extent an artefact of frequency; (2) “the law of conformity” is also an artefact of word representation models; and (3) the impact of prototypicality proposed in Dubossarsky’s earlier work is smaller.

4 Corpora: Psychology and general domain

In the current study we compare dynamics of concept breadth in two corpora: a corpus of psychology abstracts (domain-specific) and a compilation of the corpus of historical English (CoHA; Davies 2012) and the corpus of contemporary American English (CoCA; Davies 2008) texts (general domain).

4.1 Psychology corpus

The corpus comprises abstracts from journals in the field of psychology covering the period of 1930–2019 that were collected from the E-Research and the PubMed databases. In total, there are 871,340 abstracts from 875 journals resulting in 133,082,240 tokens. We only focus on abstracts since they distill the core ideas of the paper and provide a compact summary of the main contributions and findings.[2] Figure 2.1 presents the number of abstracts for each year. Due to the relatively small amount of abstracts during the first half of the 20th century, for the purpose of our experiments we only consider time periods after 1970. We also exclude two final years (2018, 2019) due to the lack of data from one of the databases.

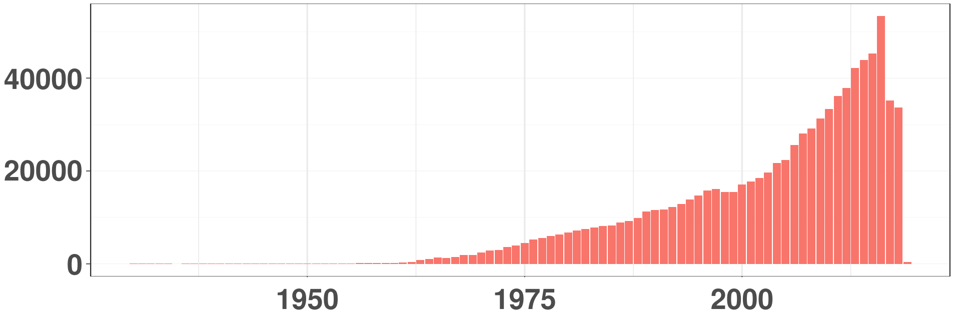


Figure 2.1: Statistics on the number of abstracts per year

[2] Restrictions related to copyright also limited our focus to abstracts.

4.2 CoCA and CoHA

The corpus of historical English (CoHA) starts in the 1810s and ends in the early 2000s, comprising 400 million words from 115,000 texts evenly sampled for each decade from fiction, magazines, newspapers, and non-fiction books.

The corpus of contemporary American English (CoCA) covers the period from 1990 till 2019 and contains about 1 billion words from 500,000 texts evenly sampled from spoken, TV shows, academic journals, fiction, magazines, newspapers, and blogs.

For the purpose of the study, we combined the two corpora leaving only the period between 1970 and 2017. We excluded blogs because of the lack of timestamps and additionally removed texts extracted from academic journals to ensure a contrast between academic and non-academic sources for our analyses.

4.3 Preprocessing steps

All corpora were preprocessed in the same way: we removed punctuation, numbers, stop-words and non-English words, did case folding and lemmatization using spaCy.[3]

The resulting corpus of psychology abstracts comprises 73,788,954 tokens from 825,628 texts. The general domain corpus has 253,597 texts with 237,205,654 tokens in total.

5 Representation of concepts

We manually associate each concept with a list of most morphologically and semantically related words. For our frequency analysis we sum the corresponding token frequencies.[4] We only consider tokens that occurred at least 50 times in each corpus. The final representation of concepts is as follows:

'Addiction': *addict*, *addiction*

'Bullying': *bully*, *bullying*

'Trauma': *trauma*, *traumatic*, *traumatize*

'Harassment': *harass*, *harassment*

'Prejudice': *prejudice*

[3] https://spacy.io/. spaCy uses a pre-trained multi-task CNN-based model that takes into account part-of-speech information (i.e. adjective *addicted* will not be transformed into *addict*).

[4] As we mentioned above, the corpus contains lemmata only.

In order to obtain concept vector representations, we follow the DISTRIBUTED DICTIONARY REPRESENTATIONS approach proposed in Garten et al. (2018) which is similar to Mendelsohn et al. (2020). More specifically, we represent each concept as a mean vector of the corresponding word vector representations (e.g., 'addiction' would be an average of vector representations of *addict* and *addition*). Unlike Mendelsohn et al. (2020) we do not assign frequency-based weights to tokens.

6 Experiments

6.1 Frequency-based analysis

For each of the five concepts we first evaluate their (unigram) frequency distribution over time. We evaluate relative frequencies by normalizing the raw counts by the total number of tokens in each year.[5]

As Figure 2.2 demonstrates, in the psychology domain all concepts demonstrate relative increase in frequency: 'trauma' exhibits the steepest slope, 'bullying' gradually raises since the 1990s, and 'harassment' has its peak in the mid-1990s. 'Addiction' and 'prejudice' present the lowest changes in relative frequency. The results obtained on CoCA/COhA (Figure 2.3) are more unsteady and labile: 'trauma' rises over time but much less rapidly compared to the psychology literature, relative frequencies of 'addiction' and 'bullying' increase over time. 'Harassment' also demonstrates the highest usage in the early 1990s while 'prejudice' slightly declines. Does the increase in the frequency of 'trauma' imply that it has broadened over time, i.e. its usages expanded to new contexts, especially in psychology literature? On the other hand, 'trauma' exhibits the highest usage among the five concepts in psychology literature, so "the law of conformity" (Hamilton et al. 2016b) would predict that it should change slower. 'Harassment' presents the lowest raw frequencies throughout most time periods but has risen in the mid-nineties. Would this imply that 'harassment' changed its meanings faster and achieved the highest breadth in the nineties?

In the next section, we adapt two diachronic variations of word2vec (Mikolov, Sutskever, et al. 2013) to quantify semantic change over time. We first train a type-based model conceptually similar to the one proposed in Mendelsohn et al. (2020). We use the type-level embeddings to obtain token-level (sentence-specific) representations which are further utilized to measure semantic breadth in each epoch.

[5] We also applied a moving average smoothing with window size of 1, i.e. $f_{1972} = (f_{1971} + f_{1972} + f_{1973})/3$.

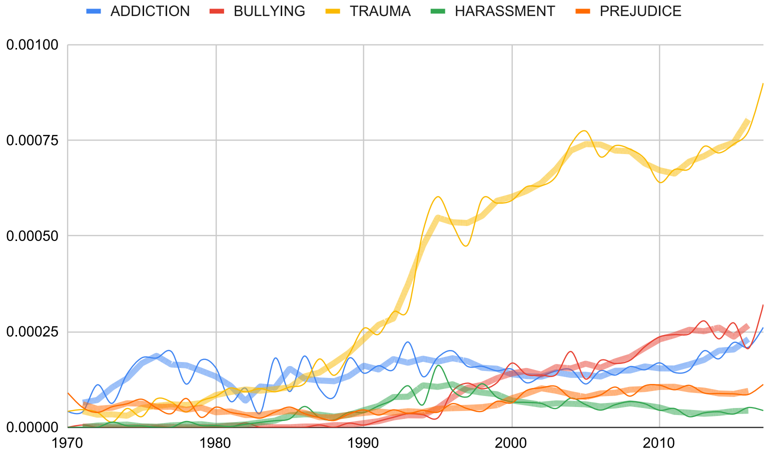


Figure 2.2: Relative concept frequencies based on abstracts from psychology journals. Bold lines correspond to moving average smoothing (window=1).

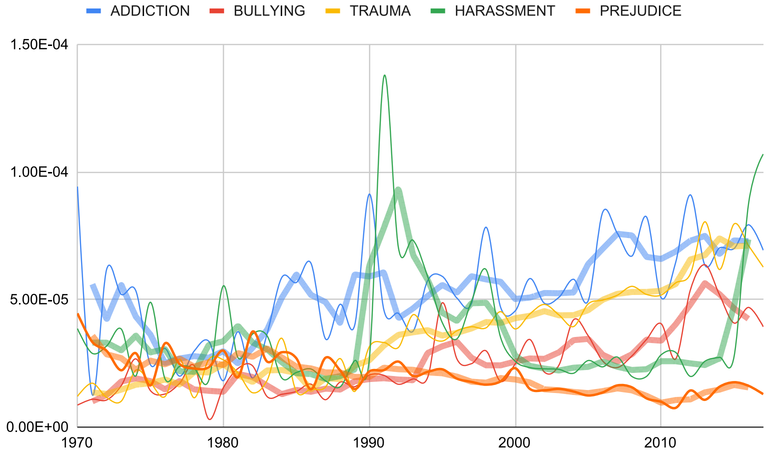


Figure 2.3: Relative concept frequencies based on general domain corpus. Bold lines correspond to moving average smoothing (window=1).

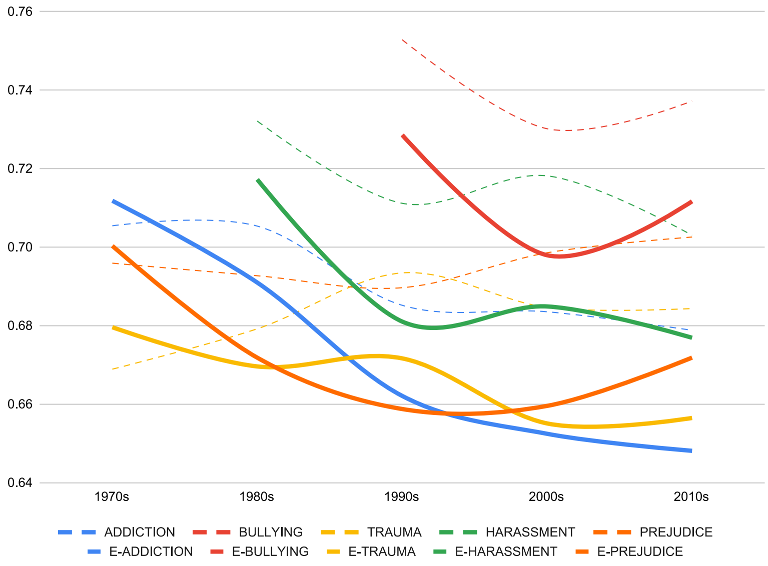


Figure 2.4: Mean cosine similarities (polynomial smoothing) over five decades (psychology abstracts corpus). Bold and dashed lines correspond to epoch-specific (e-*) and global (static) embeddings, respectively.

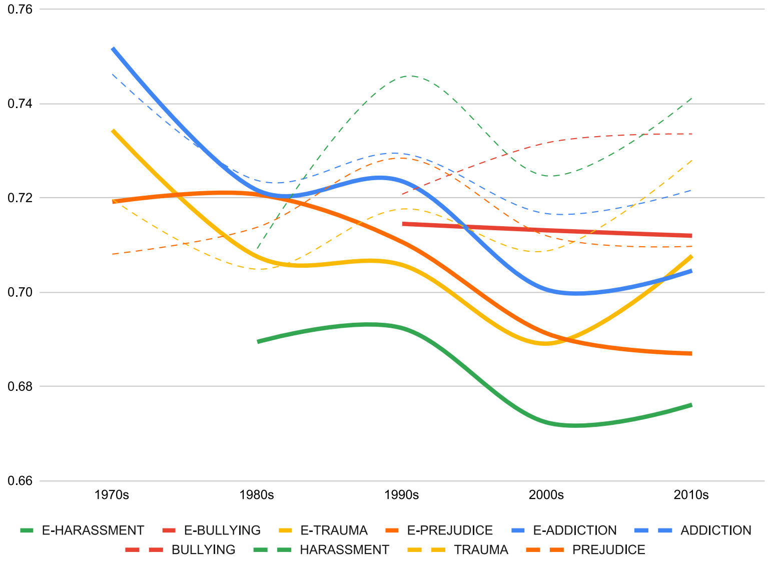


Figure 2.5: Mean cosine similarities (polynomial smoothing) over five decades (general domain corpus). Bold and dashed lines correspond to epoch-specific (e-*) and global (static) embeddings, respectively.

We then take a closer look at the type-level epoch-specific embeddings to study *how exactly* concepts changed. Such models have previously shown their utility at capturing semantic changes over time (Tahmasebi et al. 2018, Kutuzov et al. 2018) and do not require vector space alignment (which, as has been previously shown, leads to noise; Dubossarsky et al. 2019).

6.2 Diachronic word2vec

We first train a *type-level* word2vec skip-gram model.[6] In terms of hyper-parameter setting we follow that of Mendelsohn et al. (2020). Since we mainly focus on semantic changes, we set the context window size to 10 to better capture semantics and associations (Agirre et al. 2009). We also do not consider tokens that occur less than 5 times over the whole corpus. We train the model on the whole corpus for 10 iterations (obtaining *global*–static embeddings). We then use the global embeddings to initialize epoch-specific models that we continue training on each epoch's data independently for another 10 iterations. We split time periods by decades.[7]

6.2.1 Token-level embeddings

In order to obtain token-level embeddings, the resulting (global and epoch-specific) embeddings are then contextualized for each decade starting the 1970s and finishing the 2010s. This part of experiments is based on the method proposed by Sagi et al. (2009) except that we use the word2vec model (Mikolov, Sutskever, et al. 2013) rather than LSA (Landauer et al. 2013) (therefore, we refer to it as "neural parameterization of Sagi et al.'s model").

More specifically, in order to get sentence-specific vector representations for each concept in a certain decade, we randomly sample a number of its sentential occurrences[8] from the respective period, then extract contextual tokens found within the pre-set window size.[9] The final sentence-specific representation is a bag-of-words, i.e. it is an average over corresponding token representations. Following Sagi et al. (2009), in order to estimate the semantic breadth of a word,

[6] Using https://radimrehurek.com/gensim/.

[7] Due to an insufficient amount of data for earlier time periods, we train the models only on the following time frames: 1980–1989, 1990–1999, 2000–2009, 2010–2017.

[8] We set the number to 50. We use all sentential instances if the concept occurs less than 50 times during the epoch (having 20 as a minimum)

[9] We set the window size to be 3, 7, 9 tokens at each side and found that 9 provides smoother results, so we used this setting throughout.

we evaluate pair-wise cosine similarities across all the sentence-specific representations. To reduce any biases, we repeat the above sentence sampling process 10 times. The final mean values for cosine similarities for both types of models, global and epoch-specific, in the psychology and general domains are presented in Figures 2.4 and 2.5. The figures also illustrate that epoch-specific embeddings (marked as bold) provide more robust results, and we will mainly rely on them in our study.

The five concepts behave differently over time. For instance, 'trauma', although becoming frequently used in the psychology corpus, has only broadened its meaning slightly and has stayed quite a "broad" concept. In CoCA/CoHA *trauma* does not appear much before the 1990s.[10] Figure 2.5 presents two slopes; the first one can be possibly explained by the difference in its frequency distribution in CoCA and COhA, while the second one is due to its breadth changes. The notion of 'harassment', on the other hand, has the steepest slope between the 1980s and the 1990s, and then it stabilizes in its contextual usages. The highest contextual similarity in the 1980s can be partially attributed to relatively few usage instances in psychology corpus during this period. In CoCA/CoHA, frequency of 'harassment' has a drastic leap in the 1990s but, as Figure 2.5 shows; it does not affect its breadth when compared to the 1980s, although it becomes broader in the the 2000s (its usage frequency also decreases). The concept of 'bullying' has been constantly increasing in its relative usage frequency in the psychology literature, although its semantics presents a more complex pattern: it broadened from the 1990s to the 2000s, and then narrowed in the 2010s. In CoCA/CoHA the usage of 'bullying' was more stable and did not significantly change in frequency and semantics. Similarly, 'addiction' stayed within the same frequency range after the 1990s (although being much less frequent in the 1970s and the 1980s), and its breadth slightly increased since then. In the psychology domain its semantic breadth changes are more drastic: 'addiction' has been gradually becoming broader since the 1970s due to its expansion to new behavior types. Finally, 'prejudice', the concept that was not widely used before the 2000s in both corpora,[11] behaves differently in the general and psychology domains: in psychology abstracts it narrows down in the 2010s while in CoCA/CoHA it continues to expand its meaning. The results support the findings obtained for the LSA-based model in Vylomova et al. (2019). The next part of the chapter investigates how exactly the meanings changed.

[10] I.e. it is much less represented in CoHA.

[11] It appears less than 100 times a year before early 2000s.

6.2.2 Type-level embeddings

We now use the obtained epoch-specific *type-level* embeddings to run a detailed study of concept change.

Following Hamilton et al. (2016b), we consider two metrics to evaluate semantic changes over time:

1. *Semantic displacement*, which shows to what extent a concept has semantically changed during a certain time period. This is quantified as cosine distance between the word embeddings from the corresponding time periods, i.e. cos-dist($\mathbf{w}^t, \mathbf{w}^{t+\delta}$).

   Figure 2.6 shows the results of the semantic displacement evaluation and confirms our observations made earlier using the model from Sagi et al. (2009). Concepts such as ‘trauma’, ‘bullying’, ‘prejudice’ change similarly in the psychology and general domain corpora. The largest gaps are observed in the case of ‘addiction’.

2. *Pair-wise similarity time-series*, which is quantified as

   $$s^{(t)}(w_i, w_j) = \text{cos-sim}(\mathbf{w}_i^t, \mathbf{w}_j^t)$$

   and measures how cosine similarity between words w_i and w_j changes over time period $(t; t + \delta)$. For each concept we first constructed a list of words which the concept most often co-occurred with within each time period. Then we calculated cosine similarity between the concept and every word from the list for each decade. We will now discuss changes in each concept individually.

6.2.2.1 ‘Trauma’

As Figure 2.7 illustrates, ‘trauma’ has undergone more significant meaning changes in the psychology literature than in CoCA/CoHA where it preserves most associations since the 1990s. More specifically, in the psychology corpus, we observe a clear shift from *physical* to *psychological.* Although its relatedness to *severe* is still more prevalent than *mild*, they both increase their similarity to ‘trauma’ over time. In both corpora, ‘trauma’ started moving away from *childhood* in the 2000s.

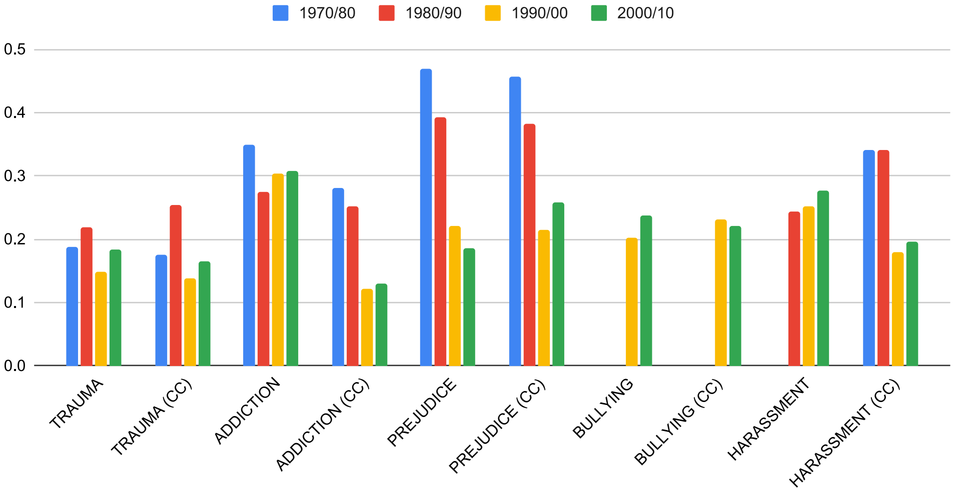
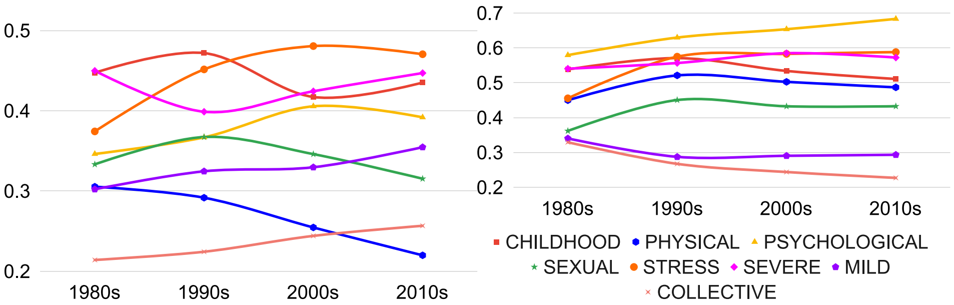


Figure 2.6: Cosine distances between decades in the psychology and CoCA/CoHA (CC) domains



(a) Psychology

(b) General Domain (CoCA+CoHA)

Figure 2.7: ‘Trauma’. Cosine similarities over four decades

Table 2.1 lists its top nearest neighbors in both corpora: ‘trauma’ stays strongly associated with ‘PTSD’. In the general domain it is associated with *horrific* and *suffer*, and its relatedness to the latter increases over time. During the 1990s–2000s ‘trauma’ becomes more *emotional* and *psychological*, which is in line with Haslam & McGrath (2020)’s findings that show changes in the relative frequency of trauma-related concepts in the massive Google Books corpus from 1960 to 2008. They found that during the 1990s the term *psychological trauma* rose most steeply.

Table 2.1: ‘Trauma’. Top-10 nearest neighbors

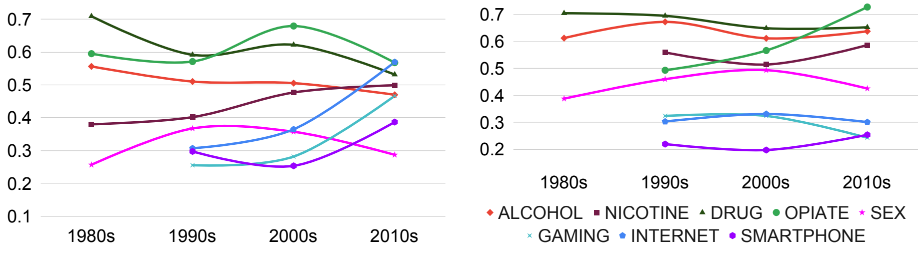
Psychology			
1980s	1990s	2000s	2010s
humbling	posttraumatic	ptsd	posttraumatic
posttraumatic	ptsd	posttraumatic	ptsd
retraumatization	survivor	traumatization	traumatization
traumatized	posttrauma	traumatized	aftermath
traumatizing	retraumatization	traumatizing	dissociative
traumatogenic	injury	desnos	peritraumatic
terrifying	traumatized	torture	traumatized
debility	atrocity	survivor	traumatically
traumatise	traumatization	dissociative	posttrauma
survivor	dissociative	posttrauma	atraumatic
traumatization	sequelae	flashback	traumatizing
unassimilable	ptsdlike	nontraumatize	pts
torture	ptds	retraumatization	refugee
traumatised	peritrauma	peritraumatic	mtbi
hypnoanalysis	desnos	nontrauma	telecommunicator
traumatolytic	psychotraumatic	lifethreat	ptss
keilson	torture	holocaust	sequelae
flashback	reexperience	nontraumatic	flashback
psychotraumatic	lasc	traumatise	postraumatic
hypnoid	traumatologist	ptes	desnos
CoCA/CoHA			
1980s	1990s	2000s	2010s
salomo	ptsd	ptsd	ptsd
posttraumatic	spiegle	posttraumatic	psychological
hyperarousal	psychological	traumatization	boehnlein
traumatization	emotional	psychological	posttraumatic
reliving	posttraumatic	emotional	hyperarousal
traumatized	horrific	horrific	suffer
indentify	psychosis	suffer	traumatization
louxes	traumatization	traumatizing	horrific
clinginess	suffer	experiencing	emotional
emotional	victim	hyperarousal	experiencing
przekop	disorder	spiegle	spiegle
brayme	sexualizing	przekop	injury
experiencing	hyperarousal	disorder	victim
ptsd	brayme	victim	traumatizing
boehnlein	abuse	painful	disorder
rohrbacher	syndrome	brayme	traumatically
spiegle	therapist	hospitalize	csf2
traumatizing	cope	severe	scurfield
csf2	boehnlein	psychiatric	przekop
traumatically	gavigan	tbi	yancosek

Table 2.2: ‘Addiction’. Top-10 nearest neighbors

Psychology			
1980s	1990s	2000s	2010s
heroin	addicted	addicted	addictive
addicted	addictive	opiate	addicted
narcotic	abuser	abuser	dependence
methadone	heroin	heroin	heroin
nonopiate	substance	addictive	mmt
illicit	dependence	drug	craving
nonaddiction	alcoholic	dependence	internet
drug	drug	substance	opiate
opiate	opiate	methadone	opioid
alcoholism	methadone	abstinence	drug
polysubstance	cocaine	cocaine	abuser
detoxification	alcoholism	detoxify	cybersex
nonaddicte	gambler	illicit	substance
alcohol	crack	detoxification	crave
abuser	abuse	abuse	problematic
detox	nonaddicte	crave	yfas
mmt	coaddict	craving	detoxification
cocaine	alcohol	abstinent	igd
polydrug	detoxification	opioid	abstinent
nonnarcotic	mmt	alcoholism	gaming
CoCA/CoHA			
1980s	1990s	2000s	2010s
drug	addicted	addicted	addicted
heroin	heroin	heroin	heroin
pcp	drug	addictive	opiate
abuser	abuser	alcoholism	opioid
methadone	alcoholic	meth	methadone
cocaine	cocaine	cocaine	alcoholism
marijuana	alcoholism	alcoholic	rehab
amphetamine	methadone	abuser	addictive
addictive	alcohol	rehab	alcoholic
alcohol	addictive	drug	suboxone
opioid	rehab	oxycontin	drug
alcoholic	oxycontin	methamphetamine	painkiller
quashen	marijuana	waismann	quashen
cannabis	abuse	medicate	overdose
opiod	henningfield	methadone	alcohol
alcoholism	buprenorphine	alcohol	vivitrol
methamphetamine	12step	quashen	acamprosate
mdma	quashen	buprenorphine	relapse
mcshin	relapse	12step	sober
addicted	addicting	relapse	oxycontin

6.2.2.2 'Addiction'

'Addiction' demonstrates a remarkable shift in the psychology literature from a substance-related concept in the 1980s to a behavior-related concept in the 2010s, but this pattern is less evident in CoCA/CoHA (see Figure 2.8 and Table 2.2). More specifically, we observe that the concept moved away from 'narcotic'-related meanings towards *gaming*, *Internet*, *cybersex*, and *smartphone*. The findings confirm earlier observations done by Vylomova et al. (2019) who used the diachronic language model from Hamilton et al. (2016b). In psychology literature, such conceptual expansion of 'addiction' had prompted and induced adaptation of a range of psychosocial treatments to be used to treat gambling, internet, and sexual addictions (Yau & Potenza 2015).



(a) Psychology

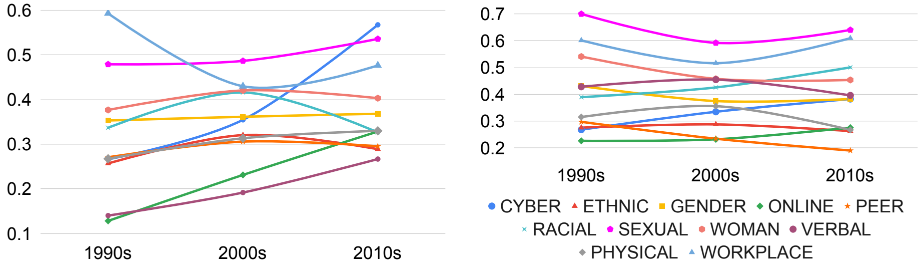
(b) General Domain (CoCA+CoHA)

Figure 2.8: 'Addiction'. Cosine similarities over four decades.

In the general domain corpora, initial associations of 'addiction' are more stable over time, and the similarity to *opiate* even increases during the last two decades. In both domains, 'addiction' becomes less associated with *abuse* and *abuser*: the similarity drops by 0.1–0.15 since the 1990s and 2000s.

6.2.2.3 'Harassment'

In both corpora, usage of 'harassment' increases in the 1990s, and the 1980s do not contain enough instances to obtain reliable embeddings. As Figure 2.9 shows, 'harassment' is highly related to *sexual* in both domains. In the psychology literature 'harassment' moves away from *workplace* towards *online* and *cyber* (increasing its relatedness to 'bullying'). In the general domain there are fewer marked changes across decades. The relationship to *online* and *cyber* is weaker than in the psychology corpus and, in contrast to that corpus, 'harassment' is more associated with *verbal* than *physical*.[12] These findings point to similarities across the corpora, but we observe a more rapidly growing preoccupation of psychology with digitally mediated forms of harassment.



(a) Psychology

(b) General Domain (CoCA+CoHA)

Figure 2.9: 'Harassment'. Cosine similarities over three decades

By looking at the nearest neighbors space shown in Table 2.3, we additionally notice substantial differences in the two domains: 'harassment' in psychology preserves its emphasis on *victimization*, the act or process of singling someone out for cruel or unfair treatment, typically through physical or emotional abuse,[13] while increases that of *perpetration*. During the 2000s–2010s it reduced its relatedness to *violence*. The general domain treats the concept of 'harassment' in a somewhat more legalistic frame, as a form of *misconduct* that is tightly associated with *allegation*, *complaint*, *accusation*, and *abuse*.

6.2.2.4 'Bullying'

Similarly, 'bullying' is markedly more victim-related in the psychology domain, having both *victimization* and *perpetration* among its top nearest neighbors in the 2000s–2010s. We additionally observe an increase in its association with 'harassment'. As shown in Figure 2.10, it becomes more associated with *workplace* while its similarity to *school* and *child* rises less steeply, consistent with its expansion into the adult realm. Similar to other concepts, we observe that bullying has expanded to cyberspace. Interestingly, its association with *cyber* accelerates upwards faster than the other concepts. As Haslam (2016) notes, referring to indirect, digitally mediated forms of aggression as "cyber-bullying" is a paradigm case of horizontal concept creep.

[12] This is probably due to 'physical' being the default characteristic of 'harassment' and usually is not explicitly marked.

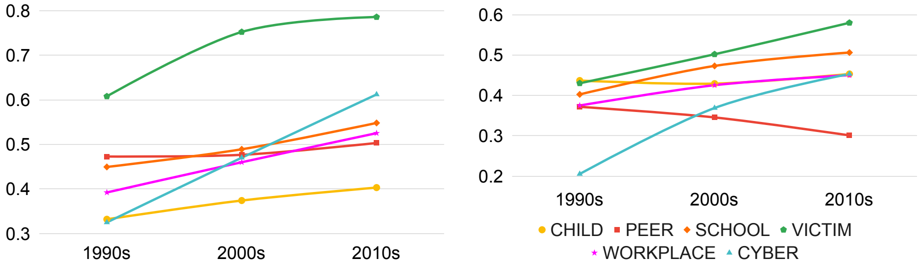
[13] The definition provided in https://dictionary.apa.org/victimization.

Table 2.3: ‘Harassment’. Top-10 nearest neighbors

	Psychology	
1990s	2000s	2010s
harasser	harasser	harasser
workplace	victimization	victimization
contrapower	contrapower	cyber
neosh	uncivil	assault
harassing	assault	bullying
uncivil	victim	perpetration
coercion	bullying	victimize
unprofessional	perpetrator	bully
assault	lsh	victim
nonharasse	victimize	perpetrate
gutek	violence	bystander
rape	perpetration	homophobic
sexualized	perpetrate	cyberbullying
nonheterosexist	rape	incivility
sexual	bully	insinuation
coercive	lgbts	heterosexist
employee	harassing	contrapower
perpetrator	cyberstalking	cyberbullye
incident	socialsexual	sexual
intragender	cyberbullye	violence
	CoCA/CoHA	
1990s	2000s	2010s
harasser	complaint	harasser
sexual	accuse	allegation
misconduct	complain	assault
complaint	intimidate	sexual
allege	assault	accusation
eeoc	intimidation	allege
accuser	allegation	complaint
sexually	harasser	workplace
allegation	misconduct	accuser
accuse	abuse	accuse
accusation	abusive	misconduct
lawsuit	threaten	rape
assault	accusation	intimidation
discrimination	renaye	lawsuit
incident	sue	harassing
abusive	discrimination	alleged
workplace	allege	defamation
abuse	sexual	abuse
rape	sexually	bullying
intimidation	mutziger	mistreatment

Table 2.4: 'Bullying'. Top-10 nearest neighbors

	Psychology	
1990s	2000s	2010s
victimisation	bullyvictim	cyberbullying
victimise	victimization	cyberbullye
cyberbullying	cyberbullying	victimization
olweus	victimisation	cyber
bullyvictim	cyberbullye	victimisation
antibullying	victimise	cyberbully
victim	olweus	perpetration
namecalling	notinvolve	antibullying
cyberbullie	victim	victimize
victimize	antibullye	victim
provictim	antibullying	harassment
cyberbullye	nonbullye	olweus
antibullye	cybervictim	antibullye
cyberbully	victimize	cyberbullie
bystanding	dipc	bystander
ringleader	cyberbullie	bystanding
bullycategorie	bullycategorie	cybervictimization
notinvolve	cyberbully	bullyvictim
nonbullye	bystander	defending
victimization	selfdestruction	kiva
	CoCA/CoHA	
1990s	2000s	2010s
aggression	bullied	cyberbullie
olweus	olweus	cyberbullying
schoolyard	coloroso	bullied
taunt	taunt	abuse
humiliate	harassment	olweus
behavior	abuser	cyberbullye
intimidate	cyberbullie	harassment
punish	himel	mutziger
abusive	intimidate	vanheest
harass	behavior	cyberbully
abuse	abuse	mehus
intimidation	13er	intimidation
aggressive	montooth	kiongozi
prosocial	milvin	nishina
taunting	nishina	taunt
13er	marrinson	insensitively
angry	aggression	zirpola
bullied	namie	sharaud
skutch	weinsheimer	fifthgrade
aggressor	vanheest	harasser



(a) Psychology

(b) General Domain (CoCA+CoHA)

Figure 2.10: ‘Bullying’. Cosine similarities over three decades

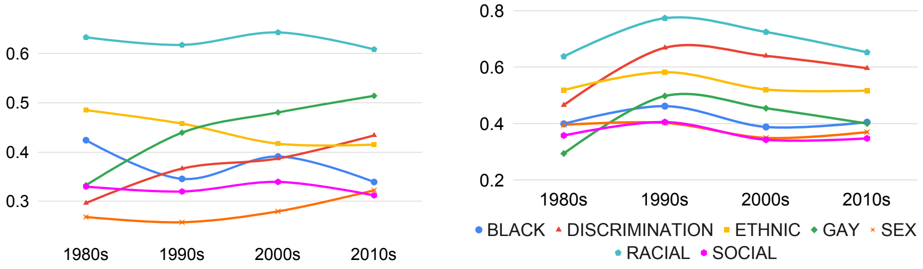
In the psychology literature, ‘bullying’ is also strongly intertwined with ‘harassment’, and both are linked to the notion of *victimization*. Arguably, this strong focus on victimization in the psychological literature, also evident in the concept of ‘harassment’, represents a preoccupation with the harm caused by bullying. The results obtained on CoCA/CoHA appear to be less congruent and more noisy, and emphasize the behaviors involved in bullying rather than the harmful impact they have on their targets. Still, it is clear that ‘bullying’ becomes more closely related to *abuse* over time in that corpus but less related to *aggression*.

6.2.2.5 ‘Prejudice’

In both corpora, but especially in psychology, ‘prejudice’ is highly associated with *racial* or *racism*, both of which are also among its nearest neighbors during all decades (see Table 2.5). In the psychology corpus, the similarity is relatively stable while in CoCA/CoHA it reduces over time. The association of ‘prejudice’ with *ethnic* and *ethnicity*, on the other hand, drops in both corpora. Dynamics of similarity with *discrimination* presents differences: it decreases in CoCA/CoHA while it rises (along with similarity to *anti-discrimination*) in psychology. The same pattern can be observed for *gay*. Interestingly, in the psychology corpus *anti-gay* and *pro-gay* are among the nearest neighbors and the similarity with *both* of them increases over time, indicating a rising attention to anti-gay prejudice within psychology over time that is not seen in the general domain. This represents a “horizontal” expansion of ‘prejudice’ in psychology beyond its earlier exclusive focus on racial animosity.[14] Analysis of nearest neighbors shows that in both domains the associations between ‘prejudice’ and *stereotyping*, *bigotry* and

[14] Among 200 nearest neighbors in each decade, the number of “anti-” and “pro-” terms is higher in psychology than in CoCA/CoHA.

belief are among the strongest and most stable over time. In the psychology literature ‘prejudice’ increases its similarity to *micro-assault* and *micro-insult* over the last decade. The growing relatedness to these forms of “micro-aggression” (Lilienfeld 2017) supports the claim that ‘prejudice’ has crept “vertically” to encompass increasingly subtle phenomena.



(a) Psychology

(b) General Domain (CoCA+CoHA)

Figure 2.11: ‘Prejudice’. Cosine similarities over four decades.

7 Conclusion

The findings of our analyses illuminate and add nuance to our understanding of concept creep within academic psychology and general domain corpora. The diachronic analysis reveals a trend for our sample of harm-related concepts to undergo semantic broadening from the 1970s to the 2010s, although the trajectories of particular concepts have been neither consistent nor linear. Since the 1990s, for example, ‘addiction’, ‘bullying’ and ‘harassment’ have broadened, as the theory of concept creep would suggest, but the breadth of ‘trauma’ and ‘prejudice’ have been relatively static. The changes are more evident in psychology literature compared to CoCA/CoHA. The analysis of semantic displacement points to a more consistent diachronic pattern: the majority of concepts changed most substantially from the 1980s to the 1990s and changed progressively less thereafter. This finding implies that societal and cultural changes occurring in the final two decades of the 20th century are likely to be especially critical for understanding concept creep. Finally, the analysis of pairwise similarities demonstrated changing patterns of co-occurrence for each concept that clarified how its meanings have shifted and expanded over four decades. During this period some concepts have acquired entirely new associations (e.g., *cyber-harassment*), some have added new semantic domains (e.g., ‘addiction’ incorporating non-ingestive

Table 2.5: ‘Prejudice’. Top-10 nearest neighbors

Psychology			
1980s	1990s	2000s	2010s
prejudiced	prejudiced	prejudiced	prejudiced
ethnocentrism	antiblack	intergroup	intergroup
xenophobia	antiforeigner	stereotyping	blatant
racial	stereotyping	blatant	stereotyping
prejudicial	stereotype	outgroup	outgroup
racism	compunction	derogated	rwa
racist	prejudicial	sdo	sdo
neuroessentialism	antigay	justif.suppression	authoritarianism
postcivil	racism	racism	derogated
ethnic	ethnopolitical	racist	antigay
ethnocentric	antifat	microinsult	justif.suppression
justif.suppression	tropp	minoritygroup	homophobia
sexblindness	antiatheist	majoritygroup	ideology
sdo	antihomosexual	antigay	rightwing
transprejudice	justif.suppression	antihomosexual	minoritygroup
intelligentsia	oldfashioned	antiblack	microassault
antiblack	neosexist	prejudicial	tropp
favoritism	intergroup	microinvalidation	antiforeigner
microinsult	multiculturalist	ingroup	microinsult
eugenics	problack	nonprejudicial	progay

CoCA/CoHA			
1980s	1990s	2000s	2010s
bigotry	racism	racism	bigotry
stereotyping	bias	bigotry	racism
racism	racial	racial	discrimination
racialist	bigotry	stereotype	stereotype
halfheartedness	discrimination	discrimination	racial
elitism	stereotype	injustice	ignorance
racial	prejudiced	racist	racist
belief	ignorance	colorism	belief
outsiderness	racist	hatred	oppression
uncomplicatedly	stereotyping	bias	bias
delegitimate	hatred	homophobia	classism
ridiculing	oppression	bigoted	misogyny
muddleheaded	injustice	belief	sexism
ethnocentrism	bigot	nonwhite	notion
factionalize	animosity	hostility	hatred
animus	homophobia	religion	discriminate
fact	bigoted	ignorance	denesh
biologism	sexism	speciesism	colorism
snideness	gender	semitism	prejudiced
multiculturalist	distrust	heterosexism	ridiculing

behaviors such as gaming and smartphone use), others have shifted emphasis (e.g., 'trauma' becoming associated less with physical injury and more with psychological stress), and yet others have come to refer to less severe phenomena (e.g., 'prejudice' becoming associated with so-called micro-aggressions). Collectively, these findings support the presence of both horizontal and vertical concept creep as proposed by Haslam (2016).

The results of the present analyses are in some respects preliminary. From a methodological standpoint, future research will need to optimize the analytic parameters employed in the approaches examined in this research and evaluate whether findings derived from these approaches converge with those using other methods for assessing semantic change. Methods must also be developed to examine horizontal and vertical concept creep separately. The methods used in the present research emphasize "horizontal" changes in the range of semantic contexts in which a concept appears, and are less capable of capturing how meanings may shift "vertically" to encompass less severe phenomena. The latter can only be inferred indirectly when concepts referring to such subtler phenomena become increasingly near semantic neighbors of the target concept.

Substantively, our findings should be replicated with additional hypothetically creeping concepts, such as 'mental illness' and 'safety'. The extent to which expansionary semantic changes are specific to harm-related concepts rather than generalized must also be studied systematically. There is scope for more focused and finely detailed analyses of semantic shifts in single concepts. Indeed, our approach offers a versatile methodology for evaluating the nature, timing, and nearest-neighbor subtleties of such shifts. Ideally, future work will explore concept creep in corpora representing other scholarly disciplines and other languages. A more fundamental challenge is to uncover the cultural factors that contribute to the semantic inflation of harm-related concepts, and to understand its societal implications.

Abbreviations

ACL	Association for Computational Linguistics
LSA	latent semantic analysis
PTSD	post-traumatic stress disorder
SVD	singular value decomposition
TF-IDF	term frequency - inverse document frequency

References

Agirre, Eneko, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca & Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL 2009*, 19–27. Boulder: ACL.

Antoniak, Maria & David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the ACL* 6. 107–119.

Belinkov, Yonatan & James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the ACL* 7. 49–72.

Bender, Emily M. & Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL 2020*, 5185–5198. Online: ACL. DOI: 10.18653/v1/2020.acl-main.463.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Bréal, Michel. 1897. *Essai de sémantique*. Paris: Hachette.

Davies, Mark. 2008. *The corpus of contemporary American English (COCA): 560 million words, 1990-present.*

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora* 7(2). 121–157.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of ACL 2019*, 457–470. Florence: ACL. DOI: 10.18653/v1/P19-1044.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer & Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of NetWordS 2015* (CEUR Workshop Proceedings 1347), 66–70. Pisa.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Fabiano, Fabian & Nick Haslam. 2020. Diagnostic inflation in the DSM: A meta-analysis of changes in the stringency of psychiatric diagnosis from DSM-III to DSM-5. *Clinical Psychology Review* 80. 101889. DOI: 10.1016/j.cpr.2020.101889.

Firth, John R. 1957. A synopsis of linguistic theory, 1930-1955. In *Studies in linguistic analysis*, 1–32. Reprinted in: Palmer, F. R. (ed.) .1968. Selected Papers of J. R. Firth 1952-59. 168-205. London: Longmans. Hoboken: Basil Blackwell.

Garten, Justin, Joe Hoover, Kate M. Johnson, Reihane Boghrati, Carol Iskiwitch & Morteza Dehghani. 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods* 50(1). 344–361.

Geeraerts, Dirk, Stefan Grondelaers & Dirk Speelman. 1999. *Convergentie en divergentie in de Nederlandse woordenschat: Een onderzoek naar kleding-en voetbaltermen.* Amsterdam: Meertens-Instituut.

Gladkova, Anna, Aleksandr Drozd & Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL student research workshop*, 8–15. San Diego.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of EMNLP 2016*, 2116–2121. Austin: ACL. DOI: 10.18653/v1/D16-1229.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2-3). 146–162.

Haslam, Nick. 2016. Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry* 27(1). 1–17.

Haslam, Nick, Brodie C. Dakin, Fabian Fabiano, Melanie J. McGrath, Joshua Rhee, Ekaterina Vylomova, Morgan Weaving & Melissa A. Wheeler. 2020. Harm inflation: Making sense of concept creep. *European Review of Social Psychology* 31(1). 254–286.

Haslam, Nick & Melanie J. McGrath. 2020. The concept creep of trauma. *Social Research: An International Quarterly* 87(3). 509–531.

Jurgens, David & Keith Stevens. 2009. Event detection in blogs using temporal random indexing. In *Proceedings of the Workshop on Events in Emerging Text Types*, 9–16. ACL.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore: ACL. DOI: 10.3115/v1/W14-2517.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*, 1384–1397. Santa Fe: ACL.

Landauer, Thomas K., Danielle S. McNamara, Simon Dennis & Walter Kintsch. 2013. *Handbook of latent semantic analysis*. Hove: Psychology Press.

Lehrer, Adrienne. 1985. The influence of semantic fields on semantic change. In *Historical semantics – historical word-formation* (Trends in Linguistics. Studies and Monographs [TiLSM] 29), 283–296. Berlin: De Gruyter Mouton. DOI: 10.1515/9783110850178.283.

Levari, David E., Daniel T. Gilbert, Timothy D. Wilson, Beau Sievers, David M. Amodio & Thalia Wheatley. 2018. Prevalence-induced concept change in human judgment. *Science* 360(6396). 1465–1467.

Lilienfeld, Scott O. 2017. Microaggressions: Strong claims, inadequate evidence. *Perspectives on Psychological Science* 12(1). 138–169.

Lukianoff, Greg & Jonathan Haidt. 2019. *The coddling of the American mind: How good intentions and bad ideas are setting up a generation for failure*. London: Penguin Books.

McGrath, Melanie J., Kathryn Randall-Dzerdz, Melissa A. Wheeler, Sean C. Murphy & Nick Haslam. 2019. Concept creepers: Individual differences in harm-related concepts and their correlates. *Personality and Individual Differences* 147. 79–84.

Mendelsohn, Julia, Yulia Tsvetkov & Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *arXiv preprint arXiv:2003.03014*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in neural information processing systems*, 3111–3119. Red Hook, NY: Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL 2013*, 746–751. Atlanta: ACL.

Pinker, Steven. 2018. *Enlightenment now: The case for reason, science, humanism, and progress*. London: Penguin.

Rogers, Anna, Olga Kovaleva & Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 104–111. Athens: ACL. https://www.aclweb.org/anthology/W09-0214.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of SemEval 2020*, 1–23. Barcelona: ACL. https://www.aclweb.org/anthology/2020.semeval-1.1.

Stern, Gustaf. 1931. *Meaning and change of meaning; with special reference to the English language*. Gothenburg: Wettergren & Kerbers.

Sturtevant, Edgar Howard. 1917. *Linguistic change: An introduction to the historical study of language*. Vol. 60. Chicago, IL: University of Chicago Press.

Sunstein, Cass R. 2018. *The power of the normal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3239204.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278*.

Vylomova, Ekaterina, Sean Murphy & Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 29–34. Florence: ACL. DOI: 10.18653/v1/W19-4704.

Vylomova, Ekaterina, Laura Rimell, Trevor Cohn & Timothy Baldwin. 2016. Take and Took, Gaggle and Goose, Book and Read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of ACL 2016*, vol. 1: Long Papers, 1671–1682. Berlin.

Weaver, Warren. 1955. Translation. *Machine translation of languages* 14(15-23). 10.

Wheeler, Melissa A., Melanie J. McGrath & Nick Haslam. 2019. Twentieth century morality: The rise and fall of moral concepts from 1900 to 2007. *PLOS ONE* 14(2). e0212267.

Winter, Bodo, Graham Thompson & Matthias Urban. 2014. Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In *Proceedings of EVOLANG 2014*, 353–360. Utrecht: World Scientific.

Xu, Yang & Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual meeting of the Cognitive Science Society, CogSci 2015*. Pasadena.

Yau, Ms Yvonne HC & Marc N. Potenza. 2015. Gambling disorder and other behavioral addictions: Recognition and treatment. *Harvard Review of Psychiatry* 23(2). 134.

Chapter 3

Computation of semantic change in scientific concepts: Case study of "circular economy"

Sampriti Mahanty, Frank Boons, Julia Handl & Riza Batista-Navarro
University of Manchester

In this chapter we aim to investigate semantic change in a scientific concept underpinned by the evolutionary framework of scientific knowledge production. The aim of this article is threefold. First is to distinguish semantic change computation in scientific concepts from that in core vocabulary and slang. Second is a multi-step analysis combining topic modelling, co-occurence networks and word embeddings, along with a control condition setup thereby presenting a pipeline to compute semantic change in a scientific concept. Third is an analysis of a popular concept in sustainability studies, i.e., "circular economy", seeking to advance research on this concept. In order to achieve our objectives, we use topic modelling to detect the point of change in a literature corpus and then we apply two approaches for detecting semantic change: co-occurence networks and word embeddings. Furthermore, we compare the concept with other related concepts in the same semantic field and use word embeddings to detect if the concept has undergone any changes relative to other concepts.

1 Introduction

Scientists contribute to the process of scientific knowledge production acting as the *central subjects* in this process. They are the entities who read the literature, perform experiments, publish the results and pass on knowledge. Textbooks and journal articles serve as *vehicles* in this process (Hull 1988). Philosophers of science have conceptualised this process of scientific knowledge production to be

Sampriti Mahanty, Frank Boons, Julia Handl & Riza Batista-Navarro. 2021. Computation of semantic change in scientific concepts: Case study of "circular economy". In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 123–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040306

evolutionary (Toulmin 1972, Hull 1988), drawing analogies with biological evolution where concepts are linguistic labels given to abstract ideas (Fodor 1975, Pinker 1994) which are framed, selected, re-conceptualised, discarded leading to a continuous evolution of the language used by researchers (Boons et al. 2017). In order to define these concepts, members of the scientific community build upon the same language (i.e., lexicon-kind terms) at the very least. Language, thus, becomes a crucial indicator to assess the shift or development in ideas (Kuhn 1990). Whilst philosophers of science have given much attention to scientific knowledge production in evolutionary terms, we intend to focus on this work from a computational perspective to understand how we can use computational methods to detect evolution in scientific knowledge production.

In the past decade an emerging research topic in the field of computational linguistics has been on the topic of semantic change computation (Tang 2018). Semantic change refers to any change in the word meaning over a period of time. Semantic change in words can sometimes happen to the extent that the modern meaning is radically different. In some cases, the semantic change that words undergo happens by means of acquiring additional meanings, rather than original meanings becoming outdated or being replaced. We find useful the definition put forward by Bloomfield (1933), where lexical semantic shifts or semantic change is defined as "innovations which change the lexical meaning rather than the grammatical function of a form". In the process of scientific knowledge production, any change in science is termed as evolution of science, thus taking *change* and *evolution* to be synonymous (Wuketits 1984, Bradie 1986). Drawing from this we set out to understand the evolution of concepts in the scientific literature through the lens of semantic change. Thus, we propose that when there is an evolution in concepts over time, it can be detected through computation of semantic change. From an empirical perspective, a key assumption is that changes in a concept's collocational patterns reflect changes in concept meaning, thus providing a usage-based account of semantics.

We begin by analysing the related work and positioning of the research in Section 2. In Section 3 we present the case study for this research. In Section 4 we discuss the methodology in detail and present the results based on our case study in Section 5. Finally, we present the discussion and conclusion in Section 6.

2 Analysis of related work and positioning of the research

There are a number of studies which have harvested the availability of huge diachronic language data to advance the research on semantic change computation (Sagi et al. 2009, Michel et al. 2011, Rohrdantz et al. 2011, Jatowt & Duh

2014, Mitra et al. 2015, Frermann & Lapata 2016, Hamilton et al. 2016, Tang et al. 2016, Dubossarsky et al. 2017). From a computational perspective, semantic change has been approached from two aspects: word-level semantic change and concept-level semantic change (Tahmasebi et al. 2018). There have been a number of studies which have focused on concept-level semantic change providing valuable insights such as the idea of concept through time (CTT, Wevers et al. 2015), parallelogram model of analogy (Orlikowski et al. 2018) and tracing concept vocabularies through a time-stamped corpus (Kenter et al. 2015, Recchia et al. 2016). However, there have been different ways in which the label of a concept has been approached in different studies. Since different domains have different interpretations of what a concept means, it should also be appreciated that any definition of a concept has a sense of arbitrariness and it is therefore desirable to study concepts with as much flexibility as possible (Fokkens et al. 2016). For instance, in studies by Kenter et al. (2015) and Orlikowski et al. (2018), there are concept terms which make up the conceptual core (core concept terms) from the rest of the vocabulary (characterising concept terms), thus distinguishing between the core and the margin of concepts. For example, for the concept of "economic efficiency", core terms might be *efficiency* and *efficient*, while characterising terms might be *robotisation*, *automatisation* or *labor productivity*. In other studies such as Wang et al. (2011), concepts not only exist in the textual information in the documents, but also refer to the quantity that a learning model is trying to predict, i.e., the variables. However, the understanding of *concept* that we use in this paper is based on some interpretations of the classical theory of concepts which treat concepts as a one-to-one correspondence with word senses (Margolis & Laurence 2019). While there are a number of schools of thought pertaining to the definition of a concept, the understanding that we adopt in this article allows us to assess the evolution of a particular concept that is used in scientific literature. Thus, going by the parsimonious understanding of a concept in our study, if we were to investigate the evolution of the "economic efficiency" concept in scientific literature, we would focus on it verbatim and possibly other forms with affixes, suffixes or short forms such as *economically efficient*, *eco-efficiency*, etc.

Hence, our computational approach is based on what has been called word-level semantic change in previous research. However, there are certain aspects which make our study distinct from previous research and bring about the novelty. The first aspect is concerned with the nature of the data and the second aspect is concerned with the rate of change.

Nature of data used: Previous studies of word-level semantic change detection are mostly based on data sources such as the Google Books Ngram corpus which is the largest text corpus used in semantic studies (Michel et al. 2011, Gulordava & Baroni 2011, Hamilton et al. 2016, Jatowt & Duh 2014, Xu & Kemp 2015, Dubossarsky et al. 2017), the Corpus of Historical American English (COHA) (Hamilton et al. 2016, Neuman et al. 2017) and the Helsinki corpus of English texts (Sagi et al. 2009). Apart from the use of such text corpora, Twitter (Mitra et al. 2015) has also been used. For the purpose of this work, our interest is in concepts used in scientific literature and we find that work on semantic change based on scientific literature is limited. Chen et al. (2018) studied semantic changes in a scientific domain by analysing the same words which have different meanings in different domains. Rudolph & Blei (2018) developed dynamic word embeddings using data from Association for Computing Machinery (ACM) abstracts and machine learning papers on the preprint database arXiv. However, most studies which use academic journal articles and abstracts as a corpus are in the biomedical domain where the classic research problem focuses on semantic relatedness and similarity between biomedical terms (Zhu et al. 2017). A recent development on computational analysis of scientific literature has been the release of SciBERT, a new resource based on contextual embeddings demonstrated to improve performance on a range of natural language processing (NLP) tasks in the scientific domain. SciBERT is a pre-trained language model based on BERT but trained on a large corpus of scientific text (Beltagy et al. 2019). In our study, we do not make use of SciBERT since: (1) we do not have access to a large enough "circular economy" corpus for training our own model, and (2) their publicly available pre-trained model was trained on a corpus where 18% of the papers were drawn from the computer science domain and 82% from the biomedical domain. Due to its narrow focus in terms of domains it is not applicable in our context. Thus, to the best of our knowledge there has been limited attention on semantic change drawn from scientific literature barring a few exceptions like the study by Dridi et al. (2019) which provides interesting insights on how to use temporal word embeddings to detect emerging scientific trends, although they do not specifically focus on semantic change. Furthermore, delving into the nature of the data, Wevers & Koolen (2020) put forward certain considerations to reflect on before training a word embedding model for computation of semantic change, i.e., (1) large enough data size spanning long time periods, (2) identification of optical character recognition (OCR) errors and spelling variations in the data, and (3) cultural and political bias in the data. These factors could affect the quality of the model being trained. However, we question if these hold true when using scientific literature as the corpus. For instance, for scientific concepts, the time period might not need to be so long as there is evidence of structural changes even within short time periods (Mahanty et al. 2019), which we believe can be detected through semantic change computation. Due to shorter time periods, the data that is available to study semantic change of concepts in scientific literature is often much smaller compared to studies using newspaper articles, Twitter data, movie reviews and books. Scientific text is more likely to be devoid of noise like OCR errors and spelling variations. Ideally, scientific literature is also devoid of any cultural and political bias since most journals have a criteria of using inclusive language in articles. Thus, the data drawn from scientific literature is of much better quality, therefore increasing the chances of obtaining a better trained model.

Rate and nature of semantic change: There are systemic irregularities in the rate of semantic change of words wherein the rate of change of some words is higher than that of others (Hamilton et al. 2016). Studies have established that the distributional properties of words implicate semantic change by showing that verbs change at a faster rate than nouns (Dubossarsky et al. 2016). Another study by Greenhill et al. (2017) is along the same line of thought and it again provides evidence on different rates of change in different aspects of language. They show that, in general, grammatical features tend to change faster and have higher amounts of conflicting signals than basic vocabulary, suggesting that subsystems of language show differing patterns of dynamics. When extending this to scientific concepts, we hypothesise that there is a difference in the rate of change between scientific concepts, core vocabulary and slang. While core vocabulary has been found to be more stable (Bengtson 2011), slang words are ephemeral (Wang 2020). Meanwhile, scientific concepts often undergo changes such as reconceptualisation, recombination and relabelling in the process of evolution (Bradie 1986), and are therefore borrowed, adapted or inflated. Thus, a blanket case of semantic change might not fit for scientific concepts.

This chapter focusses on the development of a computational pipeline for assessing conceptual evolution in the process of scientific knowledge production based on journal articles and abstracts. From a conceptual evolution perspective there are two aspects that can be investigated, i.e., firstly the evolution of a scientific concept exclusively and then the evolution of the concept with respect to other concepts in the same "semantic field". Since we understand evolution of scientific concepts by the evolution of language used by scientists, we address the first strand of our work by analysing the change in a concept's associated vocabulary. Meanwhile, for the latter strand of this work, we will investigate what is called a "semantic field", which is defined as a set of words which cover a particular semantic domain and bear structured relations with one another (Jurasky & Martin 2000). Semantic fields can be studied diachronically or synchronously. The former focusses on the origin and transformation of specific concepts while the latter deals with concepts appearing and their connection with other concepts. Thus, the understanding of a semantic field allows a better understanding of the meaning and the context of a concept with respect to other concepts. Our second goal will be to develop a computational pipeline to understand the evolution of a concept with respect to other concepts in the semantic field. In our previous work, we assessed the evolution of a concept using topic modelling (Mahanty et al. 2019). While that work enabled us to computationally detect conceptual evolution to some extent, we did not uncover the extent of change or evolution, nor determined if the change is statistically significant. These are now addressed in the work described in this chapter. Furthermore, building upon our previous work, we perform a comparison of a number of concepts within the same semantic field thereby tracing the shift in their contexts.

Our contribution in this article are the following:

a. We establish a case for semantic change in scientific concepts and its difference compared to core vocabulary and slang.

b. From a methodological perspective, we present a pipeline based on multi-angle analysis combining different methods along with a control condition setup for computation of semantic change. This can be broadly applied across disciplines to analyse a vast and expanding literature on any concept.

c. From an application perspective, this study provides insights on a very popular concept in sustainability studies.

3 Case study: The concept of "circular economy"

In the sustainability debate, the concept of "circular economy" (CE) has received immense traction amidst scholars, practitioners and policymakers in the recent years. CE refers to a system of provision in which resources are circulated between production and consumption rather than linearly transformed from production to consumption to waste. An example of CE application would be "Garment Collecting Program" of H&M, which was the result of a partnership between H&M and I:CO. This project aims at collecting and propose to their customers the used clothes at the firm's stores in three ways: "(i) rewear, i.e. clothing that can be worn again will be sold as second hand clothes; (ii) reuse, i.e. old clothes and textiles will be turned into other products, such as cleaning cloths; (iii) recycle, i.e. everything else is turned into textile fibers, or other use such as insulation" (Urbinati et al. 2017).

While the CE term was coined by Pearce & Turner (1990) it mostly underwent a dormant phase until early 2000s apart from a few mentions, for example, Cooper (1994, 1999). But from the early 2000s it has received immense attention in the academic discourse with over 1000 academic articles published in a single year. It became a popular policy agenda across European countries, China and Latin America. Another interesting aspect of CE is its relation to other concepts. The concept of CE is often studied in relation to other concepts such as "bio-economy", "green-economy" (D'Amato et al. 2017), "cradle to cradle", "industrial ecology", "closed loop supply chains", "regenerative design", "blue economy", "industrial symbiosis", "reverse logistics", "performance economy", "natural capitalism", and "biomimicry" (Geisendorf & Pietrulla 2018). Some of these concepts are termed as the antecedents to the CE concept or CE schools of thought (Homrich et al. 2018). While we assess the changes in the CE concept, it is also of interest to understand how it is evolving in relation to other concepts. It might also provide further evidence of evolution in the concept of CE, if there is any.

Thus, we use the concept of "circular economy" as a case study to fulfil two of our goals i.e.,

a. study the evolution of a concept exclusively

b. study the evolution of a concept in relation to other concepts in a semantic field.

We aim to uncover the conceptual evolution of the concept "circular economy" in academic literature through computational analysis of semantic change. We use academic articles published on CE from 2005 to 2019.

4 Methodology

In this section we discuss our methodology in detail. Figure 3.1 is a brief representation of the methodology.

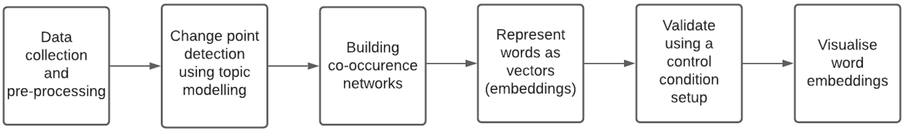

Data collection and pre-processing
Change point detection using topic modelling
Building co-occurence networks
Represent words as vectors (embeddings)
Validate using a control condition setup
Visualise word embeddings

Figure 3.1: Methodology

4.1 Data collection and preprocessing

As our first step towards our first research goal, we retrieved data in English from the Scopus[1] database using the keyword "circular economy" from the period 2005–2019. The title, abstract, author keywords, index keywords and year of publication were retrieved and stored in a comma-separated values (CSV) format. We also collected full text articles from Elsevier in XML format. This corpus consisted of 3,300 articles. In order to fulfil our second goal of analysing the conceptual evolution of CE with respect to other concepts, we created a supplementary corpus of academic abstracts on 20 other concepts that have similar or overlapping conceptualisations with CE from the period 2005–2019. The related concepts are based on literature (D'Amato et al. 2017, Geisendorf & Pietrulla 2018) and a Delphi study,[2] which we conducted with 66 academic researchers working on the CE concept. The 20 related concepts and their definitions are presented in Appendix A. A total of 61,444 abstracts was included in this supplementary corpus. In the preprocessing step we removed all instances of punctuation from the data, then the data was lower-cased and tokenised.

4.2 Change point detection

There has been a divergent policy articulation of the CE concept differentiating between its Chinese and European framings (McDowall et al. 2017). Based on such insights from the literature we hypothesise that there was a change in the concept of CE within the period of study i.e., 2005–2019. We seek to determine whether the concept changed significantly, and if yes, to estimate the point in

[1] https://www.scopus.com/home.uri

[2] A Delphi study is a forecasting process framework based on the results of multiple rounds of questionnaires sent to a panel of experts. Several rounds of questionnaires are sent out to the group of experts, and the anonymous responses are aggregated and shared with the group after each round. We asked a question to the academic researchers about what other concepts they use in their research apart from "circular economy" and they provided other overlapping concepts which they tend to use in their research articles.

time when this happened. The formulation of the task as change point detection is appropriate because even if a word might change its meaning (usage) only gradually over time, we can expect that there will be a time period when the new usage becomes much more dominant (Aminikhanghahi & Cook 2017). The identification of a change point is important because this will further form a basis for division of the corpus into "epochs" for further analysis. There have been various methods used for change point detection such as frequency analysis, syntactic analysis and distribution-based methods (Kulkarni et al. 2015). In this study we implement a distribution-based method to identify a change point. Specifically, our methodology is underpinned by topic modelling, a statistical approach whereby non-exclusive groupings of words (i.e., topics) are automatically induced based on their distribution in a corpus (Nikolenko et al. 2017). In topic modelling, every document (e.g., a journal article) is considered as consisting of a mixture of a number of topics, referred to as k. A well-known algorithm for topic modelling is Latent Dirichlet Allocation (LDA) (Blei et al. 2003), which we applied over the entire corpus using the lda (Chang 2011) and topic models (Hornik & Grün 2011) R packages. For each document d in the corpus, LDA computes the probability that d belongs to topic t, where t is any of the k topics automatically identified. The probabilities for each topic are then summed for each year based on the year of publication of each document. The sums are then visualised graphically in a stacked plot to assess the trend in the topics over the years. The determination of the number of topics and validation of the results is based on our previous work (Mahanty et al. 2019). We use the change point as reference to slice the corpus into two subsets with each subset belonging to before and after the change point. We refer to the documents in the period before the change point as the early dataset and those published in and after it as the contemporary dataset.

Once we have the proportion of the topics for each year from 2005–2019 we find the mean topic proportion for each topic in the two subsets of the corpus. Then, we run a paired t-test overall for all the topics which further assesses if the point of change is statistically significant.

4.3 Building co-occurrence networks

Co-occurrence vectors are employed in various ways to detect word level changes such as in context vectors, pointwise mutual information, temporal random indexing, or entropy in word level change detection (Tahmasebi et al. 2018). We develop co-occurrence networks based on the keywords associated with the documents. Visual keyword frequency data provides useful insights by revealing predominant trends in the keyword network of the analysed literature demonstrating a birds-eye view knowledge map (Li et al. 2019). In our study, nodes of the network correspond to the keywords (with a node for CE as the centroid), and edges indicate the co-occurrences; edge thickness represents the frequency of co-occurrence. A co-occurrence network was generated using the bibliometrix package[3] in R, for each of the epochs that was identified in the previous step. The development of the co-occurrence network is the first step to detecting the nature of changes in the concept diachronically in the two epochs. While keyword co-occurrence networks provide simple and high level information of a field, such networks are limited in their capacity because they only focus on high frequency words. Inclusion of words with lower frequencies will limit the interpretability of the network structure.

4.4 Training word embedding model

Word embeddings map high-dimension word vectors (usually produced using simple one-hot encoding representations) to low-dimension vectors to obtain global semantics (Tang 2018). Word embedding techniques that rely on the local context of the target words include Word2vec (Mikolov, Chen, et al. 2013) and Glove (Pennington et al. 2014). We trained two word embedding models using Word2Vec, one for each of the early and contemporary datasets. For this, we made use of the gensim package,[4] with a context window of four tokens and vector dimensionality of size 300 in line with the settings that have been used in previous work (Hamilton et al. 2016). The word embedding vectors were trained on each epoch and then aligned using orthogonal Procrustes transformation (Schönemann 1966) which has been applied to detect semantic change between different time periods (Hamilton et al. 2016, Dubossarsky et al. 2017, Abercrombie & Batista-Navarro 2019). We then compared word embedding vectors for the word of interest “circular economy” across the different time windows by calculating the cosine similarity between their embedding vectors calculated based on the two different periods. A lower cosine similarity between vectors is indicative of higher difference in the meaning, usage and context of a term.

[3] http://bibliometrix.org, Aria & Cuccurullo (2017).

[4] https://radimrehurek.com/gensim/models/word2vec.html, Řehůřek & Sojka (2010).

4.5 Validation of word embedding model using a control condition setup

We test the results obtained from word embeddings based on a control condition setup, given that Dubossarsky et al. (2017) identified some studies where semantic changes are largely spurious results of the word representation models on which they are based. Until we have sufficient knowledge about the interpretation of conceptual changes, inferences need to be drawn with care and verified through multiple methods (Sommerauer & Fokkens 2019). Thus, we use a control condition setup to validate the results drawn from the word embeddings.

Complementary to the genuine condition, a control condition is created where no change of meaning is expected. The underlying assumption is always that within the same dataset, the "circular economy" concept did not change its meaning. Again, unlike in the genuine condition, any changes that are observed can be attributed only to "noise" that stems from random sampling, rather than any real change in the usage or context of the concept. Therefore, any observed change in a word's meaning in the control condition can only stem from random "noise", while changes in meaning in the genuine condition are attributed to "real" semantic change in addition to "noise". In order to create a control condition we randomly sample the early dataset into two subsets (referred to as subsets A and B) and similarly created two random samples from the contemporary dataset (subsets X and Y). We then compute the mean cosine similarities between the word vectors of A and B and those of X and Y.

4.6 Comparison of CE with respect to other concepts

Word embedding models are known to successfully capture complex relationships between concepts, as manifested in the well-known word analogies task (Mikolov, Sutskever, et al. 2013), where a model aims to "solve" equations of the form "A is to B is as C is to what?" A classical example that is often used in distributional models is capturing the relation between *man* and *woman* is same as *king* and *queen* (by adding and subtracting the corresponding word vectors). Thus, it is a natural development to investigate whether changes in semantic relationships across time can also be traced by looking at the diachronic development of different distributional models (Kutuzov et al. 2018). Drawing from this idea we proceed with detecting changes in the CE concept in relation to other concepts. We construct a database of academic abstracts on 20 concepts that have similar or overlapping conceptualisations with CE from the period 2005–2019. Firstly one with all abstracts from the Scopus database using each of the 20 concepts and CE abstracts from the early period and second with all abstracts on the 20 concepts and CE abstracts from the contemporary period (identification of period based on Section 4.2). We again align the two models using orthogonal Procrustes and calculate the cosine distance of CE with the 20 other concepts in the early and contemporary period.

5 Experiments

In this section we present the results that we obtained following the methodology.

5.1 Change point detection

We summarise the results from the topic modelling in Figure 3.2. Based on the results of topic modelling (Mahanty et al. 2019) we identify two structurally different periods in the literature on CE. A structural change in the relative proportion of the identified topics was visually detected in the year 2015. In order to identify if the change in the year 2015 was significant we run a paired *t*-test based on the mean topic proportions in the early and the contemporary dataset. In the paired *t*-test the null hypothesis is that the mean difference between the two sets of observations is 0. A statistically significant *p*-value at 0.042 leads us to reject the null hypothesis and is supportive of our decision to divide the corpus into two epochs.

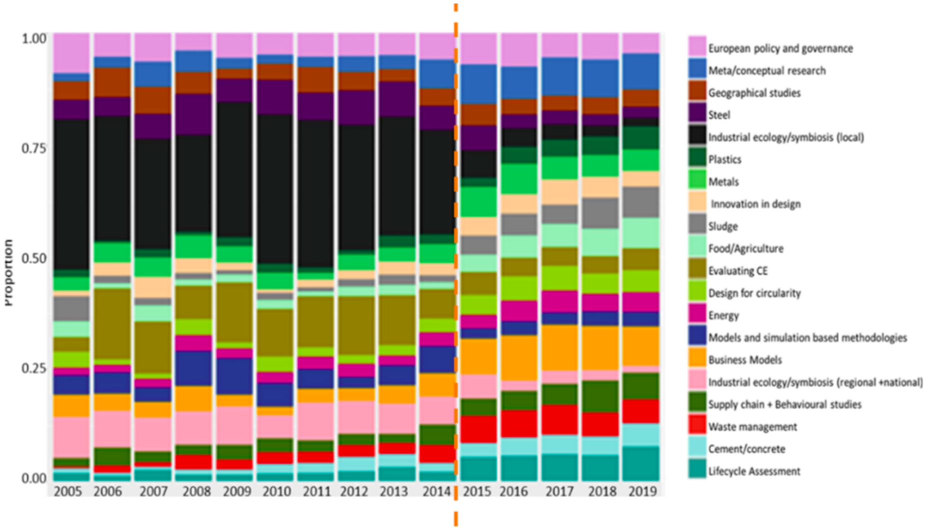


Figure 3.2: Visual representation of topic modelling and point of change

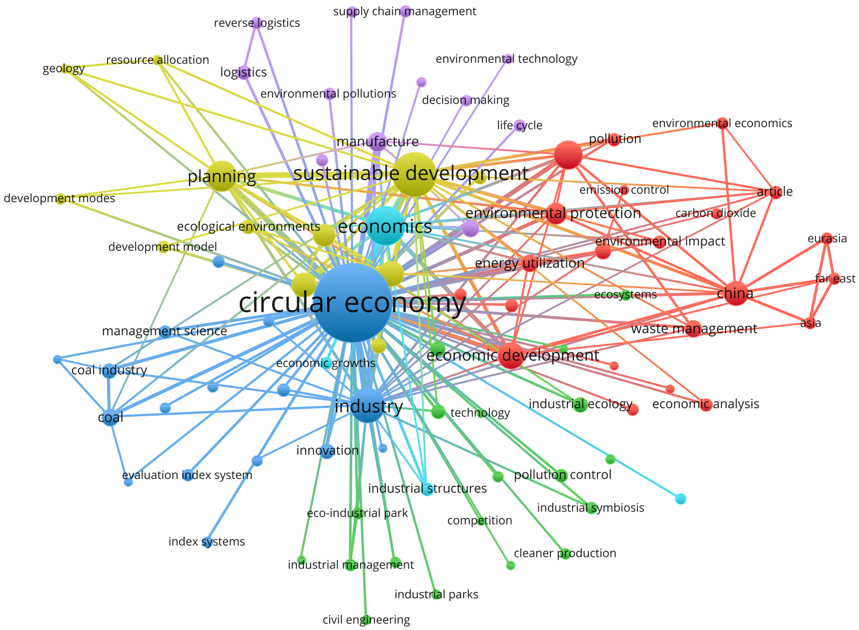


Figure 3.3: Co-occurrence network of the early dataset (2005–2014)

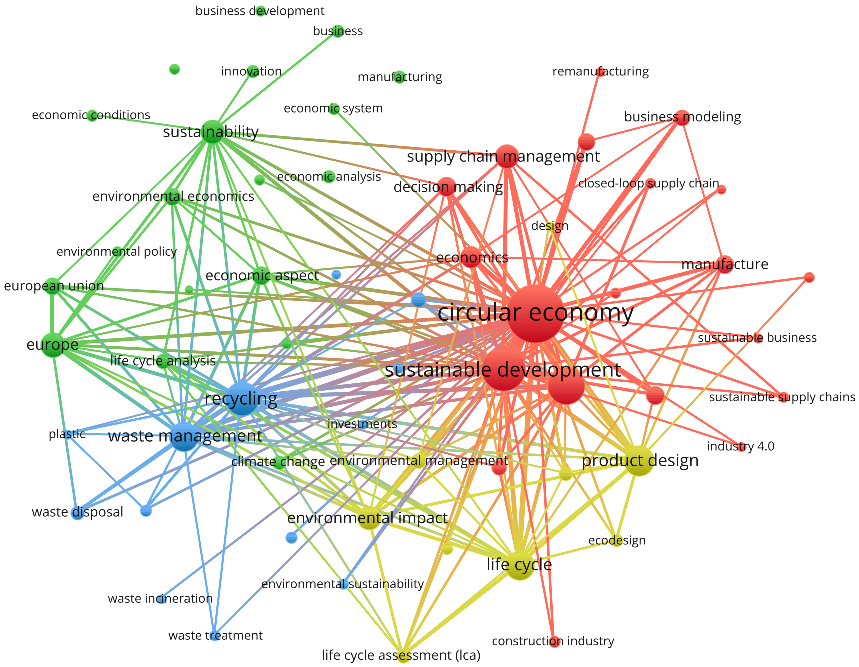


Figure 3.4: Co-occurrence network of the contemporary dataset (2015–2019)

5.2 Co-occurrence networks

On developing keyword co-occurrence networks for each of the two datasets, i.e., early and contemporary, we observed certain differences between the structures. Contemporary CE literature was found to be more strongly linked to *business models*, *supply chain*, and *product design*. Meanwhile the focus of early CE literature was more on *ecology*, *industrial economics* and *environmental management*. These observations confirm that the concept of CE has undergone some change over the years that are reflected by a shift in focus in the context of its application. We note that despite this expansion, the core meaning of the concept has not changed over time (as evidenced by the nodes that are common between the two networks, for example, *sustainable development*, *waste management*, *recycling*).

5.3 Word embeddings

After training word embeddings on each of the two datasets and aligning them using orthogonal Procrustes transformation we examined the nearest neighbours of CE (i.e., words with highest similarity to CE). We see a shift from the environmental and industrial focus to a perspective which integrates innovation with a business focus and also incorporates the social dimension of CE. The results from the word embeddings are in agreement with the results from the co-occurrence networks. The early literature primarily addressed macro-level themes in the context of environmental management and industries while the contemporary literature focuses on more micro-level interventions like business models, product design and supply chain. However, words such as *sustainability* and *sustainable development* consistently dominated the literature in both of the time periods, both of them being key to the conceptualisation of CE. The mean cosine similarity between word embedding vectors across the two time frames, i.e., early and contemporary, is only 0.195 which is quite low; this is not surprising, considering the extent of shift in the context of CE over time. We visualise the results from the word embeddings on a distributional space (Figure 3.5) using t-sne (van der Maaten & Hinton 2008) which visualises high-dimensional data by giving each datapoint a location in a two dimensional map.

5.4 Validation of results using a control condition setup

We observe the mean cosine similarity between the early and contemporary datasets is only 0.195. By using a control condition setup and creation of random

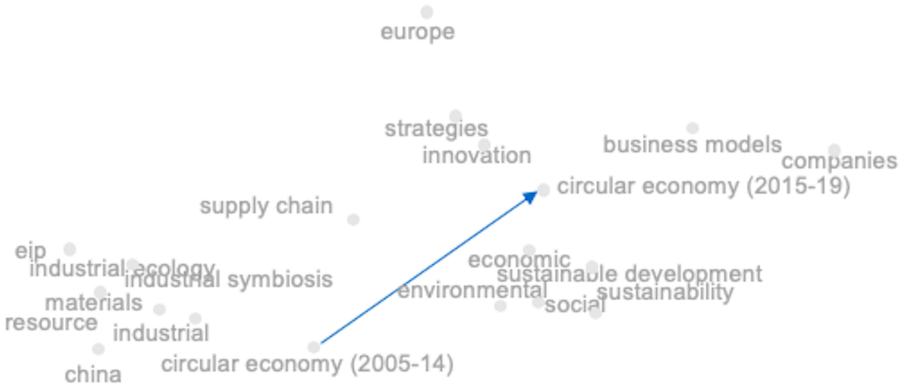


Figure 3.5: t-sne visualisation of CE based on the vectors in the early and contemporary period

subsets within the early and the contemporary period we find that the cosine similarity between the subsets drawn from the same time period was quite high, i.e., 0.62 and 0.743, for the early and the contemporary datasets, respectively. Thus, the low mean cosine similarity between early and contemporary datasets indeed indicates a change.

5.5 Comparison of CE with its overall semantic field

We compare CE with the overall semantic field and assess the relationships with the 20 concepts. In Figure 3.6 we present the semantic field in a distributional space using t-sne (van der Maaten & Hinton 2008). This is based on training word embeddings on a corpus of journal abstracts on the 20 concepts and CE. The total corpus consists of 61,444 abstracts. The individual dots represent the collocational words corresponding to each concept. The solid circles denote positions in the distributional space which are characterised by the unique contextualisation of the concepts whereas the dotted circle represents a space which constitutes an overlapping context between the concepts and depicts inter-relationships that exist between these concepts. It is interesting to note here that the "circular economy" concept seems to have an overlapping conceptualisation with most concepts. The inter-related nature of the semantic field also points towards the fact that these concepts cannot be studied or analysed in silos and researchers in these areas need to have a holistic knowledge of the associated concepts. For further analysis to detect any shift in the meaning of CE across the two time periods we divide the corpus into two parts as we did before and compute the cosine similarities between "circular economy" and each of the other concepts in the early and the contemporary period. The similarity between CE and each of the other concepts is mapped in Figure 3.7. We notice a shift in the CE concept with respect to the other concepts. Earlier the CE concept was more closely linked to "eco-civilisation" and "low-carbon economy" while in recent times it has a closer link to "sharing economy", "natural capital", and "zero waste".

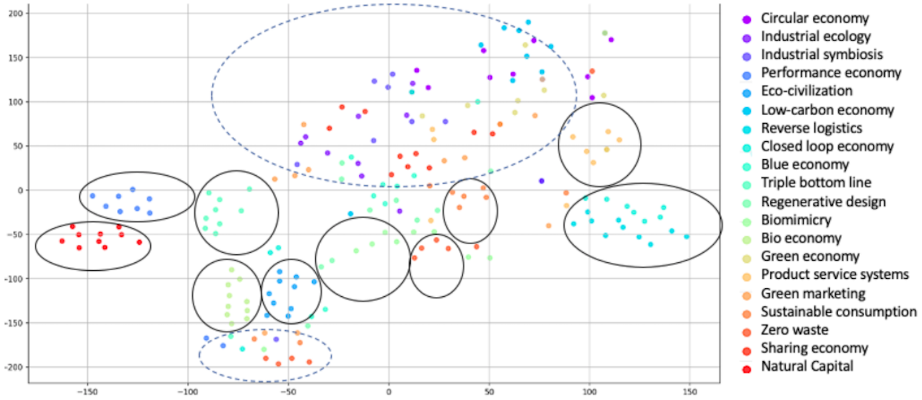


Figure 3.6: t-sne visualisation of the overall semantic field

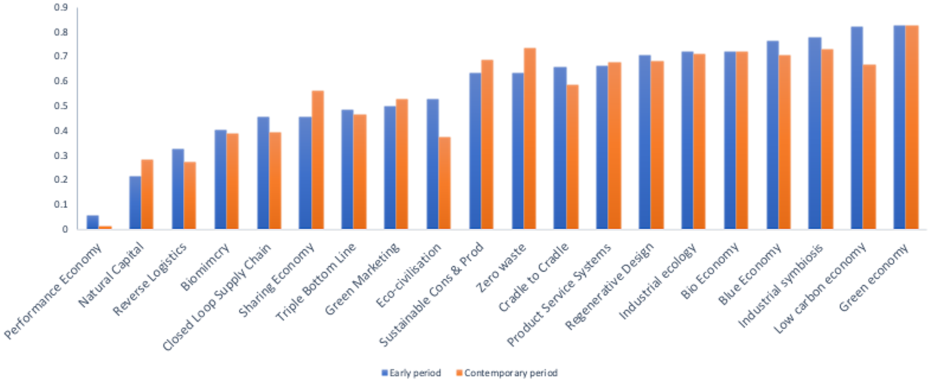


Figure 3.7: Cosine similarities of CE with other concepts in each of the two datasets

6 Discussion and conclusion

In this chapter we presented a computational approach for analysing semantic change, which is underpinned by the automated discovery of topics within a corpus of 3,300 CE academic articles in English subdivided according to their year of publication. Applying an unsupervised topic modelling method based on Latent Dirichlet Allocation (LDA) on the entire corpus, a set of topics was identified for each of the years from 2005 to 2019. A significant structural change in the relative proportion of the identified topics was detected in the year 2015. Based on this observation the corpus was divided into two broad sets, i.e., 2005–2014 (early dataset) and 2015–2019 (contemporary dataset).

To fulfill our first research objective and to detect changes in the CE concept, we compared the CE literature across these two time periods by applying on each of the data-sets two approaches – building of co-occurrence networks and training of word embeddings using "circular economy" as the primary term of interest. We then aligned the word embeddings using orthogonal Procrustes and analysed the nearest neighbours of CE and their cosine distances. In order to fulfill our second research objective to detect changes in the CE concept in relation to other concepts, we created a database of academic abstracts on 20 concepts that have similar or overlapping conceptualisations with CE from the period 2005–2019. The related concepts are based on literature and a Delphi study which we conducted with 66 academic researchers working on the CE concept. We created two datasets, firstly one with all abstracts on the 20 concepts and CE abstracts from the early period (30,762 abstracts) and second with all abstracts on the 20 concepts and CE abstracts from the contemporary period (30,682 abstracts). We again aligned the two models using orthogonal Procrustes and calculated the cosine distance of CE with the 20 other concepts in the early and contemporary period. We found that the results from co-occurrence networks and word embeddings are consistent with each other, both showing that the concept of "circular economy" has undergone semantic change. Semantic change could mean two things: either the evolution of the word usage to the point that the modern meaning is radically different or semantic change by words acquiring additional meanings rather than original meanings being outdated or being replaced. In this study we have observed the latter in the context of CE.

Specifically, our results provide computational evidence – based on three different approaches – for three main findings. Firstly, the core meaning of the concept has remained the same; this is evidenced by some common nodes in the results from the co-occurrence networks and nearest neighbours of CE based on word embeddings, such as "sustainable development", "waste management" and “recycling”. Secondly, the concept has undergone some significant expansion, where the contemporary literature on CE was more strongly linked to “business models”, “supply chain” and “product design”. In contrast, the focus of early literature was more on “ecology”, “industrial economics” and “environmental management”. Thirdly, there is a slight shift in the closeness of meaning between CE and the other concepts across the two time periods. Earlier the CE concept was more closely linked to “eco-civilisation” and “low-carbon economy” while in recent times it has a closer link to “sharing economy”, “natural capital”, and “zero waste”. The results are aligned with the history of CE where the early dataset relates to its antecedent concepts such as industrial ecology while the contemporary dataset is related to micro-level interventions for sustainable development. A further detailed analysis of the evolution of the CE concept and its relation to other concepts is beyond the scope of this paper.

From a methodological perspective, this approach could be used in assessing the evolution of concepts in academic discourse which is characterised by a vast corpus. In previous works of detecting concept change using computational methods, there have been no studies which focussed on evolution of concepts in the process of scientific knowledge production. This allows researchers to analyse large amounts of data which cannot be analysed using manual inspection. The computational methods discussed through the case study of “circular economy” could be broadly applied across disciplines. It will allow researchers to get an overview of the concept. Secondly it enables researchers to observe any high-level changes in a concept and can identify certain research directions to pursue especially in the case of vast and expanding fields. We believe that this study will be helpful for both researchers already working on related topics as well as those new to the field, for example PhD candidates who wish to quickly grasp the recent advances and history of a field and pinpoint promising research opportunities and directions. Thirdly, more often than not there are multiple concepts that exist in a particular domain and a high level analysis of the overall semantic field provides the researcher with a fair understanding of the inter-relationships between the concepts. Along with that, from a conceptual perspective an evolutionary analysis could also aid in verifying hypotheses posited by linguists, anthropologists or other researchers in a field.

Appendix A Related concepts and definitions

“industrial ecology”: Systems view which seeks to optimise the total materials cycle, from virgin materials, to finished material, to component, to product, to obsolete product, & to ultimate disposal.

"industrial symbiosis": Engaging traditionally separate industries in a collective approach to competitive advantage involving physical exchange of materials, energy, water, and by-products.

"performance economy": Represents a full shift to servitization, with revenue obtained from providing services rather than selling goods.

"eco-civilisation": Inclusion of environmental protection in the nation's economic, social, cultural, & political systems.

"reverse logistics": Process in which a manufacturer systematically accepts previously shipped products or parts from the point for consumption for possible recycling, re-manufacturing, or disposal.

"cradle to cradle": Minimizing environmental damage through sustainable production, distribution, disposal practices, & socially responsible products.

"blue economy": Optimization of natural marine resources within ecological limits, & the decoupling of environment and economy.

"triple bottom line": An accounting framework that incorporates three performance dimensions: social, environmental & financial.

"regenerative design": Principle that calls for products or services to contribute to systems that renew or replenish themselves.

"biomimmicry": Studies natures best ideas and then imitates the designs & process to solve human problems.

"bio economy": Includes all economic activities that are linked to the development & the use of biological products and processes.

"green economy": System aimed at improved "well-being & social equity, while significantly reducing environmental risks & ecological scarcities".

"product service systems": Combination of products & services in a system that provides functionality for consumers & reduces environmental impact.

"green marketing": Activities designed to generate and facilitate exchanges intended to satisfy human needs or wants, such needs and wants are satisfied without environmental impact.

"sustainable consumption and production": Use of services & products, which fulfill basic needs, bring about a better quality of life while minimizing natural resource use, toxic materials & reduce emissions thereby not jeopardising future generations.

"zero waste": An aspirational end point where all waste is reused or recycled as a resource without the need for any landfill or energy recovery.

"sharing economy": Forms of exchange facilitated through online platforms, aimed at open access to under-utilised resources through what is termed "sharing".

“natural capital”: An approach for protecting the biosphere & for improving profits and competitiveness that benefits the current and future generations.

“low-carbon economy”: Economy based on low energy consumption & low pollution.

“closed loop economy”: Used synonymously with the “circular economy”.

Acknowledgements

The authors declare no conflict of interest. Sampriti Mahanty would like to thank Alliance Manchester Business School for supporting this research work.

Abbreviations

ACL	Association for Computational Linguistics
ACM	Association for Computing Machinery
LDA	latent Dirichlet allocation
LSA	latent semantic analysis
OCR	Optical Character Recognition
XML	eXtensible Markup Language

References

Abercrombie, Gavin & Riza Theresa Batista-Navarro. 2019. Semantic change in the language of UK parliamentary debates. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 210–215.

Aminikhanghahi, Samaneh & Diane J. Cook. 2017. A survey of methods for time series change point detection. *Knowledge and Information Systems* 51(2). 339–367.

Aria, Massimo & Corrado Cuccurullo. 2017. Bibliometrix: an r-tool for comprehensive science mapping analysis. *Journal of informetrics* 11(4). 959–975.

Beltagy, Iz, Kyle Lo & Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bengtson, John D. 2011. The Basque language: History and origin. *International Journal of Modern Anthropology* 1(4). 43–59.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3. 993–1022.

Bloomfield, Leonard. 1933. *Language*. New York: Henry Holt.

Boons, Frank, Marian Chertow, Jooyoung Park, Wouter Spekkink & Han Shi. 2017. Industrial symbiosis dynamics and the problem of equivalence: Proposal for a comparative framework. *Journal of Industrial Ecology* 21(4). 938–952.

Bradie, Michael. 1986. Assessing evolutionary epistemology. *Biology and Philosophy* 1(4). 401–459.

Chang, Jonathan. 2011. *lda: Collapsed Gibbs sampling methods for topic models.*

Chen, Baitong, Ying Ding & Feicheng Ma. 2018. Semantic word shifts in a scientific domain. *Scientometrics* 117(1). 211–226.

Cooper, Tim. 1994. Beyond recycling: The longer life option. *New Economics Foundation* November. 1–21.

Cooper, Tim. 1999. Creating an economic infrastructure for sustainable product design. *Journal of Sustainable Product Design* 8. 7–17.

D'Amato, Dalia, Nils Droste, Ben Allen, Marianne Kettunen, Katja Lähtinen, Jaana Korhonen, Pekka Leskinen, Brent D. Matthies & Anne Toppinen. 2017. Green, circular, bio economy: A comparative analysis of sustainability avenues. *Journal of Cleaner Production* 168. 716–734.

Dridi, Amna, Mohamed Medhat Gaber, R. Muhammad Atif Azad & Jagdev Bhogal. 2019. Leap2trend: A temporal word embedding approach for instant detection of emerging scientific trends. *IEEE Access* 7. 176414–176428.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2016. Verbs change more than nouns: A bottom-up computational approach to semantic change. *Lingue e linguaggio* 15(1). 7–28.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Fodor, Jerry A. 1975. *The language of thought*. Vol. 5. Camrbidge, MA: Harvard university press.

Fokkens, Antske, Serge Ter Braake, Isa Maks, Davide Ceolin, et al. 2016. On the semantics of concept drift: Towards formal definitions of semantic change. In *Proceedings of Drift-a-LOD*, 247–265.

Frermann, Lea & Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the ACL* 4. 31–45. DOI: 10.1162/tacl_a_00081.

Geisendorf, Sylvie & Felicitas Pietrulla. 2018. The circular economy and circular economic concepts—a literature analysis and redefinition. *Thunderbird International Business Review* 60(5). 771–782.

Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson & Russell D. Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114(42). E8822–E8829.

Gulordava, Kristina & Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 67–71. Edinburgh: ACL. https://www.aclweb.org/anthology/W11-2508.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Homrich, Aline Sacchi, Graziela Galvao, Lorena Gamboa Abadia & Marly M. Carvalho. 2018. The circular economy umbrella: Trends and gaps on integrating pathways. *Journal of Cleaner Production* 175. 525–543.

Hornik, Kurt & Bettina Grün. 2011. Topicmodels: An R package for fitting topic models. *Journal of Statistical Software* 40(13). 1–30.

Hull, David L. 1988. *Science as a process.* Chicago: The University of Chicago.

Jatowt, Adam & Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of Joint Conference on Digital Libraries* (JCDL '14), 229–238. http://dl.acm.org/citation.cfm?id=2740769.2740809.

Jurasky, Daniel & James H. Martin. 2000. *Speech and language processing: An introduction to natural language processing*. Upper Saddle River, NJ: Prentice Hall.

Kenter, Tom, Melvin Wevers, Pim Huijnen & Maarten De Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1191–1200.

Kuhn, Thomas S. 1990. The road since structure. In *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association*, vol. 1990, 3–13. Philosophy of Science Association.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*, 1384–1397. Santa Fe: ACL.

Li, Jingwei, Pavlo D. Antonenko & Jiahui Wang. 2019. Trends and issues in multimedia learning research in 1996–2016: A bibliometric analysis. *Educational Research Review* 28. 100282.

Mahanty, Sampriti, Frank Boons, Julia Handl & Riza Batista-Navarro. 2019. Studying the evolution of the 'circular economy'concept using topic modelling. In *International Conference on Intelligent Data Engineering and Automated Learning*, 259–270. Springer.

Margolis, Eric & Stephen Laurence. 2019. Concepts. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Summer 2019. Stanford: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/concepts.

McDowall, Will, Yong Geng, Beijia Huang, Eva Barteková, Raimund Bleischwitz, Serdar Türkeli, René Kemp & Teresa Doménech. 2017. Circular economy policies in China and Europe. *Journal of Industrial Ecology* 21(3). 651–661.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331. 176–182.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint 1301.3781*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in neural information processing systems*, 3111–3119. Red Hook, NY: Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal & Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21(5). 773–798.

Neuman, Yair, Harvey Hames & Yochai Cohen. 2017. An information-based procedure for measuring semantic change in historical data. *Measurement* 105. 130–135.

Nikolenko, Sergey I., Sergei Koltcov & Olessia Koltsova. 2017. Topic modelling for qualitative studies. *Journal of Information Science* 43(1). 88–102.

Orlikowski, Matthias, Matthias Hartung & Philipp Cimiano. 2018. Learning diachronic analogies to analyze concept change. In *Proceedings of the second joint sighum workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, 1–11. Santa Fe: ACL. https://www.aclweb.org/anthology/W18-4501.

Pearce, David W. & R. Kerry Turner. 1990. *Economics of natural resources and the environment*. Baltimore, MD: JHU press.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*, 1532–1543. Doha: ACL. DOI: 10.3115/v1/D14-1162.

Pinker, Steven. 1994. *The language instinct. The new science of language and mind*. London: Penguin.

Recchia, Gabriel, Ewan Jones, Paul Nulty, John Regan & Peter de Bolla. 2016. Tracing shifting conceptual vocabularies through time. In *European knowledge acquisition workshop*, 19–28. Springer.

Řehůřek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, 45–50. http://is.muni.cz/publication/884893/en. Valletta: ELRA.

Rohrdantz, Christian, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim & Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of HLT 2011: Short papers*, 305–310. ACL.

Rudolph, Maja R. & David M. Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of WWW 2018*, 1003–1011. ACM. DOI: 10.1145/3178876.3185999.

Sagi, Eyal, Stefan Kaufmann & Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, 104–111. Athens: ACL. https://www.aclweb.org/anthology/W09-0214.

Schönemann, Peter H. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31(1). 1–10.

Sommerauer, Pia & Antske Fokkens. 2019. Conceptual change and distributional semantic models: An exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 223–233.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278*.

Tang, Xuri. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering* 24(5). 649–676.

Tang, Xuri, Weiguang Qu & Xiaohe Chen. 2016. Semantic change computation: A successive approach. *World Wide Web* 19(3). 375–415. DOI: 10.1007/s11280-014-0316-y.

Toulmin, Stephen. 1972. *Human understanding: The collective use and evolution of concepts*. English. Princeton, N.J.: Princeton University Press.

Urbinati, Andrea, Davide Chiaroni & Vittorio Chiesa. 2017. Towards a new taxonomy of circular economy business models. *Journal of Cleaner Production* 168. 487–498.

van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9. 2579–2605.

Wang, Ling. 2020. Analysis of the characteristics and translation skills of American slang in the Big Bang theory. *Theory and Practice in Language Studies* 10(10). 1248–1253.

Wang, Shenghui, Stefan Schlobach & Michel Klein. 2011. Concept drift and how to identify it. *Journal of Web Semantics* 9(3). 247–265.

Wevers, Melvin, Tom Kenter & Pim Huijnen. 2015. Concepts through time: Tracing concepts in Dutch newspaper discourse (1890–1990) using word embeddings. *Digital Humanities* 2015.

Wevers, Melvin & Marijn Koolen. 2020. Digital begriffsgeschichte: Tracing semantic change using word embeddings. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53(4). 226–243. DOI: 10.1080/01615440.2020.1760157.

Wuketits, Franz M. 1984. Evolutionary epistemology—a challenge to science and philosophy. In Franz M. Wuketits (ed.), *Concepts and approaches in evolutionary epistemology*, 1–33. Dordrecht: Springer.

Xu, Yang & Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual meeting of the Cognitive Science Society, CogSci 2015.* Pasadena.

Zhu, Yongjun, Erjia Yan & Fei Wang. 2017. Semantic relatedness and similarity of biomedical terms: Examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Medical Informatics and Decision Making* 17(1). 1–8.

Chapter 4

Semantic change in Swedish – from a lexicographic perspective

Stellan Petersson & Emma Sköldberg
University of Gothenburg

In this chapter, we examine semantic change in the general vocabulary of present-day Swedish and its lexicographic description. We discuss the question of whether automatic and semi-automatic methods of computational linguistics are relevant to lexicography and conclude that such methods can facilitate, formalize, and sharpen lexicographic investigations of semantic change.

1 Introduction

Several efforts have been made to automate or semi-automate parts of the process of dictionary compilation, including the building of headword lists and identification of collocations (Cook et al. 2014). Automatic methods for finding linguistic examples have also been developed (see, e.g., Kilgarriff et al. 2008, Pilán 2016). Furthermore, there are computational linguistic studies that examine semantic changes in large text corpora (e.g. Cavallin 2012, Cook et al. 2013, Nimb et al. 2020). A central aim of studies of this kind is to make lexicographic work more efficient; another, related aim is to introduce more systematicity into the process of dictionary construction. The results of studies like these are, of course, relevant to practical dictionary editing. In the ongoing work on the forthcoming second edition of the dictionary *Svensk ordbok utgiven av Svenska Akademien* (‘The contemporary dictionary of the Swedish Academy’, henceforth “SO”), semantic changes on the lexical level are important. However, the editorial group (of which the authors are members) currently lacks formal, computational methods for discovering semantic changes on the lexical level.

Stellan Petersson & Emma Sköldberg. 2021. Semantic change in Swedish – from a lexicographic perspective. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 149–167. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040308

The main purpose of this chapter is to discuss lexicographical problems that are associated with a number of Swedish examples, where each type of example represents an area of research on lexical semantic change. Furthermore, the chapter addresses how computational linguistics and language technology can facilitate lexicographical research in this area.

In Section 2, we begin by providing a general characterisation of our lexicographical framework, focusing on the database from which different editions of SO are based. We then proceed to a discussion of a number of Swedish examples, in Section 3, and explore several lexicographical issues that are relevant to the database and its development. Finally, in Section 4, we turn to the interface between lexicography and computational linguistics and provide some remarks on research in this area relevant to the work on SO.

2 SO and the lexical database

The focus in this chapter is on SO, a definition dictionary containing about 65,000 headwords describing the general vocabulary of modern Swedish. The emphasis in the dictionary is on the meanings and uses of the words. SO, which is corpus-based, is primarily aimed at users with Swedish as their mother tongue, but also at learners with good knowledge of Swedish. The first edition of SO was published in book form in 2009. It is now also available in a digital format, as an app and through the Swedish Academy's dictionary portal, svenska.se. The second edition of SO is scheduled to be published in 2021, but only in digital format.

SO is a subset of a very extensive lexical database (currently including approximately 200,000 headwords) which has been under continual development at the University of Gothenburg (GU) since the 1970s. According to a collaboration agreement between the Swedish Academy and GU from 2010, GU will further develop and maintain the extensive database until 2060. The database, which a research group at the Department of Swedish is responsible for, consists of new words, word forms, and word connections continuously incorporated into the database. The publisher's aim is to release a new edition of SO at least every ten years. For each edition, SO will provide as complete information as possible about every important word and expression in Swedish. This information includes the word's spelling, pronunciation, inflection, style, emotive charge, and meaning. Each word entry will be illustrated with language examples but also with examples of phraseology and constructions.

The 2009 edition of SO has several merits, but because large parts of the dictionary articles were compiled in the 1980s, the dictionary can be improved upon and modernized in several ways. A fundamental part of the revision work with SO has been to review the headword list. Since the semantic description of the headwords is so central to the dictionary, an important part of the revision work has been to examine whether the meanings of the headwords have changed since the first edition of the dictionary. In conjunction with this work, what these changes consist of has been subject to analysis.

It is evident that the information about semantic change and the new meanings that are present in the database must be compatible with the description model that is already established in SO. According to Svensén (2009: 211–212), the polysemy structure of the words in a dictionary can be described linearly, i.e., as a number of discrete units arranged in a sequence. This observation is primarily valid for monolingual Swedish dictionaries like *Natur och Kulturs stora svenska ordbok* (Köhler & Messius 2006) and *Bonniers svenska ordbok* (Sjögren & Györki 2010). The same is valid for many dictionaries of English, like the *Longman dictionary of contemporary English* and the *Merriam-Webster dictionary*. However, the polysemy structure of a word can also be described in terms of a limited number of main/core senses, to which groups of subsenses/shades are associated (i.e., in an hierarchical order of senses). This is the approach that has been adopted in SO, where the way in which the subsense(s) are related to the main sense is also explicitly specified. This principle is also valid in *Den Danske Ordbog* ("DDO"), the most comprehensive monolingual dictionary of contemporary Danish. To illustrate how these two different principles work in these dictionaries, consider the example *ansiktslyftning* 'face-lift'.[1] First of all, the word refers to surgery, but it is also used metaphorically to refer to repairs that make a building, for example, look newer or better. When the polysemy structure is described linearly in the dictionary, these two meanings are listed as 'meaning 1' and 'meaning 2'. However, when the polysemy structure of the dictionary is hierarchical, the two meanings are listed as 'meaning 1' and 'meaning 1a', because the second meaning is considered to be a metaphorical subsense of the main sense ('meaning 1').

In the hierarchical variant, the relationship between the senses is typically categorized in terms of meaning extension, meaning specialization, metaphorical (figurative) use, etc. An example of meaning extension can be seen in the meaning of the noun *visitkort* 'visiting card'/'business card'. *Visitkort* originally only referred to a kind of concrete paper card with printed information (name, company name, address, etc.). Nowadays, we can find digital visiting cards as well. The process of meaning specialization can be exemplified by the noun *nedtrappning* cf. 'tapering', referring to a gradual decrease of something. The noun

[1] All English translations in this chapter are by the authors.

has a special meaning in a medical context: it refers to the process of gradually lessening or reducing the use of a medicine, etc. Finally, an example of a semantic change which is based on metaphorical (figurative) use of a existing word is the already-mentioned *ansiktslyftning* 'face-lift'. (See e.g. Malmgren 1988: 181, for a discussion of meaning relations in the lexical database, and the classic Waldron 1967 for important background information in this area.)

Svensén (2009) points out that an important point in arranging the meanings is to determine how far a subsense of a meaning should be allowed to depart from the main sense before the lexicographer has to consider establishing a new and independent sense. Svensén (2009: 212, 363) also concludes that once the division into senses has been implemented, the order in which the senses are to be presented must be determined. Traditionally, dictionaries have applied a historical order, starting with the oldest sense and ending with the most recent (see, e.g., *Svenska Akademiens ordbok*, 'The Swedish Academy dictionary', henceforth "SAOB"). However, an arrangement of this kind has disadvantages, for example, this approach is not suitable for the majority of users, since they might give up before they find the meaning they were looking for.

In summary, a central goal for the editorial team of SO is both to provide a correct description of contemporary Swedish and to show the relationships between different senses. At the same time, the team aims at compiling a lexicographic resource that is an understandable and useful tool for its intended users in different user situations (including the reception and production of Swedish).

3 Examples from the lexicographer's shop floor

In this section, we examine a number of Swedish words that represent well-known types of semantic change, which have been explored in previous research on developments in the Swedish vocabulary during the last decades: changes in concepts and their reference (discussed in Svensén 2009), emotive change (related to, for example, "feminist language change"; see Wojahn 2015: 35–52), changes in constructional behaviour (Malmgren 2003), and, finally, grammaticalisation and pragmaticalisation (see, e.g., Rosenkvist & Skärlund 2011). Our main contribution is to highlight and discuss several lexicographic problems associated with these types of change.

3.1 Concepts, distinctive features, and prototypical meanings

Definitions in dictionaries are associated with word forms or lemmas. According to a common assumption, definitions are assumed to correspond to concepts in the minds of language users. As normally understood in the field of lexicography, an intensional definition of a concept, the standard format for definitions in general-language dictionaries, consists of a superordinate concept (genus proximum) of the concept to be defined and at least one distinctive feature (differentia specifica) specific to the concept in question (Svensén 2009: 218–221). The distinctive features specify in which respects the concept to be defined differs from other concepts that are related in the same way (subordinated) to the genus proximum. For example, consider the concepts 'quadrilateral', 'rhombus', and 'rectangle'. An intensional definition of 'rhombus' states that genus proximum is 'quadrilateral' and adds one or several features that distinguishes it from the concept of rectangle, e.g., that all sides have equal length. Importantly, the number of features has to be adjusted so that the definition does not become too narrow or too broad (see also Atkins & Rundell 2008: 414–417).

In SO, nouns and verbs are defined in this way. For example, the main sense of *örn* 'eagle' is *typ av stor rovfågel med långa breda vingar, kraftig näbb och grova klor…* ('type of large bird of prey with long broad wings, strong beak and robust claws'). Another example is *bryta*, literally 'break' (c.f. *broken*), where one of the main senses is *tala med främmande uttalsmönster…* ('speak with a foreign accent').

However, the genus-and-differentia model is sometimes unworkable, since large areas of the lexicon do not fit this taxonomic model. Furthermore, the goal of identifying the necessary and sufficient distinctive features of a lexical unit is questionable (see, e.g., Atkins & Rundell 2008: 416). According to prototype theory, it is impossible to determine which distinctive features are both necessary and sufficient in defining a certain category, since the borderlines between categories are fuzzy (Svensén 2009: 224, see also Rosch & Mervis 1975). Consequently, the lexicographer may aim for a typification in the meaning description by analyzing many individual instances of words in a corpora, instead of trying to isolate necessary and sufficient conditions. The dictionary user will then see the definition that is normally or typically the intended one (see Atkins & Rundell 2008: 418 with references and Svensén 2009: 222–223).

But what happens when, for example, a category of concrete objects, which a lexical item refers to, radically changes over time? How does such a change affect the definition of the word in the dictionary? And what happens when a lexical item referring to a certain kind of state or condition appears in new contexts?

The first case can be illustrated by the noun *bil* 'car'. According to SAOB, the word was established more than a hundred years ago and is currently in use today. Whilst there has been a remarkable technological development of cars, has the meaning of the word changed? In the following, we focus on the lexicographical consequences of the technical development that has taken place with respect to these vehicles. In SO the main sense of the word is described in the following way: *motordrivet, (vanligen) fyrhjuligt fordon med plats för ett litet antal personer och vanligen främst avsett för persontransport* ('motorized, usually four-wheeled vehicle with room for a small number of people and usually primarily intended for passenger transport'). Although the type of fuel has varied over time and that there exist self-driving cars nowadays, the SO definition, with its superordinate concept and distinctive features, is so general that it points out both older and younger car models. In other words, it still points out a typical car and hence the definition does not need to be updated. However, by using language examples like *familjebil* 'family car', *småbil* 'small car', *elbil* 'electric car', *hybridbil* 'hydrid car', *en bensinsnål bil* 'a petrol-efficient car', *en fyrhjulsdriven bil* 'a four-wheel drive car' and *en förarlös bil* 'a self-driving car', the lexicographer indicates that there is a certain range within the concept and that the referents of the noun can be quite different.

The second case, when a lexical item referring to a certain kind of state or condition appears in new contexts, is illustrated by the abstract noun *nollvision* 'vision zero'. When this word was introduced in the 1990s, it was the name of a government safety project, which had the aim that no one should be killed or seriously injured as a result of a traffic accident in Sweden. The name has become a common noun in Swedish, and in SO (2009) it is defined in the following way: *vision som går ut på att ingen ska dödas eller skadas allvarligt i trafiken* ('vision that no one should be killed or seriously injured in traffic'). However, nowadays the noun also appears in other contexts (cf. the process of generalisation in the well known Waldron 1967). The word may also concern societal aims with regards to the number of suicides and cases of domestic violence. In this particular case, the lexicographers of the second edition have to decide whether the definition from 2009 should be (1) reformulated or (2) complemented. The definition from the first edition can, of course, be revised so that it becomes more general. Since SO is a synchronic dictionary, the sense development of the word during the last decades does not have to be shown. Alternatively, the definition from 2009 can be regarded as the main sense, and a subsense referring to the same kind of safety policy in other contexts could be added. According to the usage in modern corpora, the traffic context is the most recurrent and typical. For this reason, it is likely that the editorial team of the second edition of SO will choose the second option.

3.2 The development of a semantic field: Computers and information technology

There is a clear connection between society and vocabulary, as the seminal works Ullmann (1962) and Waldron (1967) point out. The vocabulary is the most flexible part of a language and new words are introduced hand in hand with new things, new ideas. At the same time, words disappear as the social reality that the word refers to changes. *Nyordsboken* (Moberg 2000: 11–12), which elaborates on the ideas of Ullman and Waldron, presents the following five different ways in which the Swedish vocabulary expands:

1. Already existing words or expressions are given a new meaning or a new area of use.

2. Existing words are combined into a new compound word or phrase.

3. A derivation suffix is added to an already existing word.

4. Existing words or expressions are abbreviated.

5. Words or expressions are borrowed from other languages.

Points 1 and 5 are the most relevant in the present context. Frequently, new words are created through a combination of these ways. Most words that have acquired a new meaning in addition to an already existing meaning are semantic loans (borrowed meanings). In such cases, a word that already exists in Swedish acquires a new meaning through the influence of a foreign word.

Several examples of semantic loans can be found in the field of computers and information technology. Swedish words like *virus* ‘virus’, *mus* ‘mouse’, *ikon* ‘icon’, *mapp* ‘folder’, *portal* ‘portal’, and *surfa* ‘surf’ were already established when their new computer-related meanings were borrowed into Swedish. Hence, these words became (if not already) polysemous.

A quick look in the first edition of SO shows that the SO (2009) editorial team have chosen to treat these new meanings in slightly different ways. Consequently, one can discern certain inconsistencies in how the different words were treated. Despite the fact that all the new meanings are metaphorical (figurative), in some cases the new meanings have formed the basis for a special main sense (as in the case of *mus* and *portal*) and sometimes for a sub-sense to the senses already established (as in the case of *virus* and *surfa*). When defining a new sense, the lexicographers have regarded the new sense as semantically remote and separated from the meanings already described (cf. Svensén 2009 in Section 2).

Some verbs with computer-related meanings that are not registered in SO (2009), but which will most likely be added to the forthcoming second edition, are *importera* 'import', *exportera* 'export', and *strömma* 'stream'. In the first two cases, we observe metaphorical but also specialized uses of the words. The new meaning of *strömma* is semantically related to the basic meaning of the same verb, but while the traditional use of the verb is intransitive, the new one can also be used transitively (cf. SOMETHING *strömmar* 'streams' vs. SOMEONE *strömmar* 'streams' SOMETHING). This aspect is also taken into account in the analysis of the verb and may affect how the new meaning is treated in the dictionary.

A slightly different kind of word that has received a marked increase in use, not least through social media, is the noun *hatare* 'hater'. According to SAOB, the word has been used since 1541 in Swedish texts. Without doubt, the traditional meaning and the new use have many semantic features in common, but the context in which the new meaning appears should of course be included in a description of how the word is normally or typically used today.

Recently, new uses of the Swedish verbs *posta* and *texta* (in the senses 'to publish on the internet' and 'to send an sms') have been noticed (cf. the traditional meanings 'to post a letter' and 'to write in block letters'). There is no doubt that these new uses of the words occur in young people's spoken language. The question is, however, whether these uses are sufficiently established. This can be determined by searches in different corpora.

Whether the above observations reflect a change in the meaning of existing words can also be discussed. The semantic difference between the more established and the new uses of *texta* is so striking that it can be argued that *texta*, in the sense 'send an sms', is simply a new word; it is a homonym to *texta* 'to write in block letters' that has appeared in Swedish.[2] According to Ullmann (1962: 59), homonymy refers to the fact that two synchronically different words have the same surface form and polysemy to the fact that one word has two or more different senses (see Atkins & Rundell 2008: 280–282 for a discussion of polysemy and homonymy in English dictionaries).

Another example from the same subject area is *troll*. The Old Swedish word *troll* 'ugly and supernatural being with a tail; usually perceived as hostile to humans' and the English equivalent 'troll, elf' have been used since the end of the 13th century. Since at least 2009, however, we find an identical word in Swedish with the meaning 'Internet troll'. In terms of surface form, the Swedish nouns *troll*

[2] Note that the Swedish *posta* 'publish online, typically on a social media website' is often pronounced in a semi-English fashion (but this is not visible in writing). This also affects the lexicographic classification of the word.

‘ugly being with tail’ and *troll* ‘on the Internet’ coincide completely. The words have the same pronunciation and inflection. Furthermore, they have specific semantic aspects in common. For this reason, one could argue that *troll* is polysemous and that the latter use has evolved from the former. The new *troll* is then an example of semantic change. However, the two nouns have completely different origins in that *troll* ‘being’ is derived from the ancient Swedish *trul, trol*, whereas *troll* ‘on the Internet’ comes from the English verb and noun *troll* with origins in ‘to fish by trolling’ (https://www.merriam-webster.com/dictionary/trolling). For this reason, it is more reasonable to consider *troll/troll* as homonyms instead of arguing that the word is polysemous. However, as the widely-consulted Lyons (1968: 406) states:

> The distinction between homonymy and multiple meaning is, in the last resort, indeterminate and arbitrary. Ultimately, it rests upon either the lexicographer’s judgement about the plausibility of the assumed ‘extension’ of meaning and some historical evidence that the particular ‘extension’ has in fact taken place.

Lyons (1968: 406) also points out that “the arbitrariness of the distinction between homonymy and multiple meaning is reflected in the discrepancies in classification between different dictionaries” (cf. Svensén 2009 in Section 2).

3.3 Emotive meaning

The cases above concern cognitive (or denotative) meaning. There are, however, also dimensions of emotive (or connotative) meaning in language (see Svensén 2009: 214 for a discussion of different notions of meaning). A natural starting point for lexicographers is Stevenson’s famous discussion of this notion:

> The emotive meaning of a word is a tendency of a word, arising through the history of its usage, to produce (result from) affective responses in people. It is the immediate aura of feeling which hovers about a word. Such tendencies to produce affective responses cling to words very tenaciously. It would be difficult, for instance, to express merriment by using the interjection “alas”. Because of the persistence of such affective tendencies (among other reasons) it becomes feasible to classify them as “meanings”. (Stevenson 1937: 23).

Stevenson claims that the emotive meaning of a given word arises “through the history of its usage”. The emotive aspects of meaning are closely intertwined with attitudes towards the concept referred to; it is therefore expected that the emotive meaning can change over time. Consider the word *democracy*, discussed by Stevenson. Today, the word has a positive emotive meaning, but one can imagine that democratic forms of government might fall out of popularity: the word would then keep its cognitive meaning but change its emotive (Stevenson 1944: 72).

In present-day Swedish, consider the examples *lapp* 'Lapp', *eskimå* 'Eskimo', and *indian* 'American Indian'. These words were previously stylistically neutral in Swedish. Today, however, they are not appropriate in newspaper texts and similar genres. The perception of these words has clearly changed in recent history. In all these three cases, the viewpoint of the ethnic group is today a relevant factor for most speakers and writers of public discourse. Notwithstanding this concern, it might not be obvious what the view of the relevant group is. In the case of *lapp*, the denoted ethnic group (the Sami of northern Scandinavia) perceive the expression as strongly derogatory,[3] but the viewpoints of the groups referred to by the Swedish expressions *eskimå* and *indian* are not as obvious. In the latter two cases there exist disagreements about proper labels within the groups (Bird 1999; see the entry on *eskimåer* in the *Nationalencyklopedin*[4]). Furthermore, such disagreements are often embedded in larger complex cultural and social debates in North America and South America (in the case of the English *Indian*, which is, of course, clearly related to the Swedish word but, nevertheless, a different one), and in Denmark and Greenland (in the case of Swedish *eskimå*, closely related to the Danish *eskimo*). These debates and discourses influence the public discourse in Sweden, and the Swedish language, but the distance from the debates increases both the variation in emotive charge across the population of Swedish speakers and the felt complexity, or perhaps unclarity, in relation to the emotive meaning of these words. In the forthcoming second edition of SO, *lapp* has the usage marker *starkt nedsättande* ('strongly derogatory'). *Eskimå* and *indian* have the usage marker *kan uppfattas som nedsättande* ('can be perceived as derogatory'), which highlights the variation and complexity of the words' emotive charge.

Other words, not associated with ethnicity, have also changed in emotive meaning. Consider *bög* 'male homosexual' and *flata* 'female homosexual', in the semantic field of sexual orientation. These words used to be clearly derogatory; now they have been partly reclaimed by the LGBTQI-community and can be used with a neutral emotive meaning. The meanings are, however, context dependent in these cases, and the conditions for application are quite complex: the emotive

[3] See e.g. Sametinget (2017: 5).

[4] https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/eskim%C3%A5er

charge depends on, for instance, tone of voice, discourse topic, and the identity of the speaker (cf. Petersson & Sköldberg 2020).

It is difficult to determine when a word has changed its emotive meaning. The issue is pressing, since such modifications can occur rapidly and words quickly can become controversial in the public sphere. Debates about a word in newspapers and on social media can be indicators, but the intuitions of the lexicographer play a role as well. In Section 4, we discuss whether (and how) automatic methods can be of help in detecting changes in emotive meanings.

3.4 Constructional behaviour

Another topic in the field of semantic change concerns different kinds of constructions and word combinations. It is well-known that a language user's mastery of the syntagmatic properties of lexical items (including relevant collocations) has major consequences for whether their language is perceived as idiomatic or not. Research has also shown that even advanced learners have difficulties with the use of different kinds of conventionalized expressions in their second language (see, e.g., Nation 2013: 479 with references). SO aims to account for this theme, in addition to providing a complete description of headwords and their syntagmatic properties.

A word that has evolved considerably in recent decades is the reflexive verb *gifta sig* 'get married'. The main sense of the word, 'enter into marriage', has been established since the 14th century and, for centuries, the constructions for the word were SOMEONE *gifter sig* 'gets married' (with SOMEONE) or SOME PEOPLE *gifter sig*. Changes in society have had consequences for this verb, however. Since same-sex marriages regularly take place in society, the gender of the referents of SOMEONE and SOME PEOPLE, the subject and the object, has undergone change. However, because the definition of the verb is *ingå äktenskap* 'to marry' which includes both classes of marriage, it has not been revised in the (forthcoming) second edition of the dictionary. The relatively new situation, where the referents of the subject and object can be of the same gender can be illustrated by the following language example in a dictionary article: *hon har gift sig med sin fästmö* 'she has married her fiancée$_{\text{[F]}}$'.

Since the 1990s, there is also a metaphorical use of the same verb. In corpora, we can find examples such as *låt såsen dra i några timmar innan servering så att smakerna hinner gifta sig* 'let the sauce soak/draw for a few hours before serving so that the flavors have time to get married' with the constructions SOMETHING *gifter sig* (with SOMETHING) and SOME THINGS *gifter sig*. This sense was well-established when the first edition of SO was compiled, but, due to a lack of more advanced tools, it was not noticed by the lexicographers. By using corpus tools (for instance the ones provided by the research unit Språkbanken Text, see *Korp* (Borin et al. 2012)) it is now much easier for lexicographers to register this type of semantic variation.

The current search interface of Korp (Borin et al. 2012) includes three independent ways of viewing the results of a search. These are (i) the KWIC concordance view, (ii) the statistics view, and (iii) the Word Picture view.

According to Atkins & Rundell (2008), a KWIC concordance is a basic corpus lexicography tool. Right-sorted and left-sorted concordances often give a "powerful, visual representation of a word's recurrent patterns – in a way that is impossible to ignore or overlook" (Atkins & Rundell 2008: 105). By using the statistics view, it is, for example, relatively simple to compare the frequency of different spelling variants of a word.

Finally, to further develop the treatment of collocations in a dictionary, the Word Picture view is very useful. The Word Picture view in Språkbanken, which is based on the association measure lexicographer's mutual information, gives an overview of selected syntactical environments of a word (i.e., typical verbs, prepositions, pre-modifiers, and post-modifiers) (see Borin et al. 2012: 476; cf. the word sketches generated by Sketch Engine, a corpus tool presented in Kilgarriff et al. 2004). Consequently, by using Word Picture, the lexicographer can provide a more comprehensive description of the semantics of headwords (like *gifta sig*) and their phraseological behaviour.

Another kind of example concerns collocations including the word *gripa* 'profoundly touch, move, affect'. In his 2003 diachronic study, Malmgren states that the verb has become more frequent in abstract transitional phrases, such as *gripas av förtjusning/misströstan/raseri/svårmod* 'be affected by delight/despair/rage/discouragement'. At the same time, the use of the verb *falla* 'fall, yield, give way to' in corresponding phrases has become less frequent during the 20th century. Older uses such as *falla i frestelse/förtjusning/misströstan/raseri* 'give way to temptation/delight/despair/rage' are now perceived as obsolete (Malmgren 2003: 140–141). We thus observe that some collocation verbs have become obsolete and are replaced by other verbs. In other words, some verbs (like *gripa*) demonstrate expanded or extended combinatorial properties while other verbs (like *falla*) are subject to reduced and more limited combinatorial properties. Tools like Word Picture can be useful for future studies in this area.

3.5 Pragmaticalisation

Beeching (2010) explores pragmaticalisation, a process closely related to grammaticalisation (Traugott & Dasher 2001). Pragmaticalisation takes place when a content expression develops into a pragmatic marker, in contrast to grammaticalisation, which concerns the development into purely grammatical functions. Beeching focuses on the English *effectively* and the French *effectivement* and shows that the expressions have developed from the shared meanings 'efficaciously' and 'in fact' to different pragmatic meanings in the two languages: in French 'that is so' (used as an answer to a question), in English 'contrary to experience' or a purely hedging meaning (expressing uncertainty about the speaker's assertion). The explanation is related to Traugott & Dasher (2001), where a theory of grammaticalisation in terms of conversational implicatures is put forward (see the classic in pragmatics Grice 1975); in short, repeated implicatures can over time become integrated parts of semantic meanings. In the English case, the meaning 'in fact' and 'contrary to experience' invite the inference that the speaker is not making a certain assertion.

Related examples in Swedish are *typ* 'type' and *exakt* 'exactly'. These expressions have developed from content words with clear cognitive (denotative) meanings to pragmatic markers (discourse particles). Rosenkvist & Skärlund (2011) shows how *typ* develops from a noun (*en envis typ* 'a stubborn kind of fellow') via a two-word preposition *av typ* (*en båt av typ lyxjakt* 'a boat of a luxury type'), to a preposition (*Han gillar musik typ Dylan* 'He likes music of Dylan's kind') and then, in recent history, to a pragmatic marker used for hedging (*Välkommen till England, typ* 'Welcome to England, or whatever'). *Exakt* is used as an interjection, affirming previous statements (parallel to the pragmatic function of the English *exactly*).

In the dictionary, all of these uses should be described. However, SO is faced with a number of challenges, with regards to the examples reported on here. First, the different uses of *typ* and *exakt* are related, but the standard labels for meaning relations, which concern mechanisms of metaphor, generalisation, and similar types of change, are not suitable in this context (see Section 2). A new set of labels for pragmatic meaning relations is called for. Second, the structure of the headwords and main senses in SO treats the different uses of *typ* and *exakt* as different headwords, since they differ in word class. It is debatable whether a strict adherence to principles concerning the structure of the dictionary is relevant here; perhaps a more user-friendly approach would be to list all uses of *typ*, and all uses of *exakt*, under the same headword (see Section 2).

4 Discussion

In this chapter, we have discussed semantic changes in Swedish words from a lexicographic perspective. The starting point for the reasoning has been, first and foremost, the work conducted by the editorial team for the forthcoming second edition of SO, a comprehensive synchronic dictionary with emphasis on the semantics of the lexical units.

In the chapter, we have discussed a number of different kinds of semantic changes. The change can consist of a certain word taking new meaning (e.g. *nollvision* 'vision zero'). Sometimes, one is able to identify a similar development in a group of words belonging to the same semantic field (e.g., *virus* 'virus', *strömma* 'stream', etc.). The semantic change can also consist of a word being associated with more negative emotive meaning in public discourse (e.g., *indian* 'American Indian'). Such changes might happen fairly quickly. Furthermore, semantic change can consist of a changed constructional behaviour of a word. The referents of the subject and the object of a verb might shift (as in *gifta sig* 'get married') and the tendency of a word to be included in collocations may increase or be reduced (*gripa* 'profoundly touch, move, affect', and *falla* 'fall, yield, give way to'). A word may also, by pragmaticalisation, lose its lexical meaning and become a function word (e.g., *typ* 'kind of'). It is also the case that a semantic change of a word may be relatively established among some language users and within a certain kind of language (e.g., spoken youth language) but it may be relatively unknown among other groups of language users. This is the case for *posta* 'post (on blogs etc.)'. Finally, it may be the case that what one initially might have thought was a new meaning of an established word is, in fact, a new word, i.e., a homonym (as demonstrated in our discussion of *troll* 'Internet troll'). In summary, the phenomenon of semantic change, in Swedish and other languages, is multifaceted and diverse. But regardless of the type of change one examines, all types are relevant to lexicographers because these changes should lead to revisions of dictionary articles.

The development of more formal, computational tools for discovering semantic changes is most welcome. In closing, we share some thoughts on how such methods can assist us in achieving our aims with SO.

First, given the fact that SO aims to reflect general vocabulary, we are primarily interested in changes in the general vocabulary of Swedish and not in developments of meaning in technical language. Notwithstanding this stated aim, we would like to obtain more information about the differences in the semantics and the usage of Swedish words in newspaper language and in social media, for example. Our major area of interest determines which materials we should examine. However, we fear that the Swedish corpora available at the moment are too limited, and we thus propose that the existing Swedish corpora, especially with regards to the inclusion of newer texts, need to be radically improved.

Second, lexicographers have traditionally focused on written language. The main focus of lexicographers has been, and remains, on the description of established (lexicalized) changes with a relatively good spread in different corpora. However, we would also like to have more data on spoken language. Although perhaps expensive and practically difficult, a point on our wish list is a searchable corpus of authentic spoken dialogues.

Third, several semantic changes can be identified by use of Word Picture in Korp and similar technologies. It is clear that such technology is significantly helpful for observing metaphorical and metonymical changes, and for specifications and generalisations as well. However, it should be noted that any analysis of the data that is provided by different corpus tools is highly dependent on the lexicographer's linguistic intuitions and experience in the field.

Fourth, it seems to be the case that certain semantic changes cannot be identified by technologies such as Word Picture or other automatically generated information about linguistic contexts. Emotive meanings are especially difficult to identify using such techniques. In these cases, debates about words, in public discourse and social media, play a pivotal role, but the lexicographer's linguistic intuitions are crucial as well. Language technology can provide useful information about the genres, where controversial words are discussed and written about. For instance, if words related to minority groups are increasingly used on social media, that would be useful for us to know.

Fifth, similar difficulties arise in cases of pragmaticalisation and the related process of grammaticalisation. It is unclear to us how Word Picture or a similar tool would be of use in these regards. However, we could start from the problems and questions of lexicography and list a number of items that we would like to keep track of with automatic methods. This list would then include cases like *typ* and *exakt*, where it is clear that pragmaticalisation has taken place.

Finally, we claim that computational methods may be of use in studies of collocations. By examining changes over time in the narrower context of words, one can register new meanings and uses of verbs like *gifta sig* 'get married', *gripa* 'profoundly touch, move, affect' and *falla* 'fall, yield, give way to'. See, e.g., the study of variations in bigrams over time in Nimb et al. (2020), where a method of updating headwords in DDO with new semantic information is investigated. Their study, which combines corpus statistics with manual annotations, is based on "the hypothesis that the variation in bigrams over time in a corpus might indicate changes in the meaning of one of the words" (Nimb et al. 2020: 112). Furthermore, the fact that verbs such as *exportera*, *importera*, and *strömma* and the noun *hatare* are now used in computer contexts should be discernible if one compares broader contexts in corpora reflecting language from different periods of time. This can be related to Cook et al. (2013), who, based on an automatic word sense induction system, compare three sentence contexts of target words in two corpora representing different language periods, and evaluate whether there are any differences in usage of the target words.

We also suggest that language technology may be relevant to contrastive studies. Researchers studying different languages can benefit from each other's work. For example, in the field of computers and information technology, one can see clear parallels between the development of different words in English and in Swedish (see Section 3.2). Slightly simplified, if, for example, the English word *virus* begins to be used metaphorically about computers in English, it is not surprising if the same development in the corresponding Swedish noun is observed. In this context, computational methods would be most welcome.

From our perspective of lexicography, the point of computational methods is to provide sharper tools and allow for a more precise and formal methodology. In practice, language technologists might provide lexicographers with candidate lists of lexical items that seem to have undergone a semantic change. These data sets could then be assessed by lexicographers (see the methods in e.g. Cook et al. 2013 and Nimb et al. 2020). The methods for detecting semantic change would then, hopefully, become more precise. The production of dictionaries would also become more systematic and less reliant on the subjective judgment of individual lexicographers (see Cook et al. 2013: 50).

A pertinent issue in this discussion is deciding on which semantic changes should be prioritized. All of the cases discussed in this chapter are of interest, from a linguistic point of view. But for the dictionary user, especially second-language learners, the most important examples are, perhaps, the examples with negative, or unclear, emotive (connotative) meanings. Therefore, automatic methods that could help us improve the lexicographic quality with regards to the emotive aspects of words, and the changes in emotive meaning, would be most welcome.

Abbreviations

SO	The contemporary dictionary of the Swedish Academy
GU	The University of Gothenburg
SAOB	The Swedish Academy dictionary
KWIC	Keyword-in-context

References

Atkins, Sue & Michael Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Beeching, Kate. 2010. Semantic change: Evidence from false friends. *Languages in Contrast* 10(2). 139–165.

Bird, Michael Yellow. 1999. What we want to be called: Indigenous people's perspectives on racial and ethnic identity labels. *American Indian Quarterly* 23(2). 1–21.

Borin, Lars, Markus Forsberg & Johan Roxendal. 2012. Korp: The corpus infrastructure of Språkbanken. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of LREC 2012*, 474–478. Istanbul: ELRA. https://spraakbanken.gu.se/korp.

Cavallin, Karin. 2012. Exploring semantic change with lexical sets. In *Proceedings of the XV EURALEX international congress*, 1018–1022. Oslo: EURALEX.

Cook, Paul, Jey Han Lau, M. Rundell, Diana McCarthy & Timothy Baldwin. 2013. A lexicographic appraisal of an automatic approach for detecting new word-senses. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.), *Proceeedings of eLex 2013*, 49–65. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Cook, Paul, Jey Han Lau, Michael Rundell, Diana McCarthy & Timothy Baldwin. 2014. Applying a word-sense induction system to the automatic extraction of diverse dictionary examples. In *Proceedings of the 16th EURALEX International Congress*, 319–328. Ljubljana: Ljubljana University Press.

Det Danske Sprog- og Litteraturselskab (ed.). 1918–. *Den Danske Ordbog*. https://ordnet.dk/ddo.

Grice, Herbert Paul. 1975. Logic and conversation. In P. Cole & J. Morgan (eds.), *Syntax and semantics*, vol. 3, Speech acts, 41–58. New York: Academic Press.

Kilgarriff, Adam, Husák Miloš, Katy McAdam, Michael Rundell & Rychlý Pavel. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (eds.), *Proceedings of the XIII EURALEX international congress*, 425–433. Barcelona: Universitat Pompeu Fabra.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz & David Tugwell. 2004. The sketch engine. In *Proceedings of the 11th EURALEX International Congress*, 105–115. Lorient: Université de Bretagne Sud.

Köhler, Per Olof & Ulla Messius. 2006. *Natur och Kulturs stora svenska ordbok*. Christian Mattsson (ed.). Stockholm: Natur och Kultur.

Lyons, John. 1968. *Introduction to theoretical lingustics.* Cambridge: Cambridge University Press.

Malmgren, Sven-Göran. 1988. On regular polysemy in Swedish. In *Studies in computer-aided lexicology* (Data Linguistica 18), 179–200. Stockholm: Almqvist & Wiksell.

Malmgren, Sven-Göran. 2003. Begå eller ta självmord? Om svenska kollokationer och deras förändringsbenägenhet [Commit or take suicide: On Swedish collocations and their tendency to change]. *Språk & Stil* 13. 123–168.

Merriam-Webster (ed.). 2019. *Merriam-Webster dictionary.* Version 5.2.0. Springfield: Merriam-Webster.

Moberg, Lena. 2000. *Nyordsboken.* Stockholm: Svenska språknämnden.

Nation, Paul. 2013. *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nimb, Sanni, Nicolai Hartvig Sørensen & Henrik Lorentzen. 2020. Updating the dictionary: Semantic change identification based on change in bigrams over time. *Slovenščina 2.0* 8(2). 112–138.

Pearson Longman (ed.). 2009. *Longman dictionary of contemporary English.* 5th. Harlow, UK.

Petersson, Stellan & Emma Sköldberg. 2020. To discriminate between discrimination and inclusion: A lexicographer's dilemma. In *EURALEX XIXI book of proceedings Vol. 1.* 381–386. Alexandroupolis: EURAC.

Pilán, Ildikó. 2016. Detecting context dependence in exercise item candidates selected from corpora. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 151–161. San Diego: ACL.

Rosch, Eleanor & Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7. 573–605.

Rosenkvist, Henrik & Sanna Skärlund. 2011. Grammatikalisering i nutid – utvecklingen av *typ* fram till 2009 [Present-day grammaticalization – the development of *typ* until 2009]. *Språk & Stil* 21. 5–25.

Sametinget. 2017. *Svenska Sametingets kommentar till Sveriges 22:a och 23:a periodiska rapport till Kommittén för avskaffande av rasdiskriminering* Comments by the Swedish Sami Parliament regarding Sweden's 22nd and 23rd periodic reports to the Committee on the Elimination of Racial Discrimination. Kiruna/-Giron: Sametinget. https://www.sametinget.se/117555.

Sjögren, Peter A. & Iréne Györki. 2010. *Bonniers svenska ordbok.* Stockholm: Bonnier.

Stevenson, Charles Leslie. 1937. The emotive meaning of ethical terms. *Mind* 181. 14–31.

Stevenson, Charles Leslie. 1944. *Ethics and language*. New Haven: Yale University Press.

Svensén, Bo. 2009. *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

Svenska Akademien. 2009. *Svensk ordbok utgiven av Svenska Akademien*. https://svenska.se.

Svenska Akademien. 1898–. *Svenska Akademiens ordbok*. https://svenska.se/.

Traugott, Elizabeth Closs & Richard B. Dasher. 2001. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Ullmann, Stephen. 1962. *Semantics: An introduction to the science of meaning*. Oxford: Basil Blackwell.

Waldron, Ronald A. 1967. *Sense and sense development*. London: Andre Deutch Limited.

Wojahn, Daniel. 2015. *Språkaktivism: Diskussioner om feministiska språkförändringar i Sverige från 1960-talet till 2015* [Language activism. Discussions on feminist language change in Sweden from the 1960s until 2015] (Studier utgivna av Institutionen för nordiska språk vid Uppsala universitet 92). Uppsala: Institutionen för nordiska språk, Uppsala universitet.

Chapter 5

Historical changes in semantic weights of sub-word units

Yang Xu 徐炀 & Zheng-sheng Zhang 张正生
San Diego State University

In this chapter, we present a computational study on how the weight of sub-word units in determining word meanings evolves chronologically in different languages. Sub-word units, e.g., morphemes, syllables etc., play variable roles in determining word meanings. Some morphemes in English have standalone lexical meanings (e.g., the root) while others function more as morpho-syntactic markers (e.g., the bound morphemes such as *-ness* etc.) The semantic weight of sub-word units changes over time; for instance, some ancient characters in Chinese or ancient prefixes in English no longer carry clear semantic meanings. The goal of this chapter is to characterize such a change with computational methods. The semantic weight of sub-word units can be captured by word embedding models (and their variants).

We present results from two substudies. In Study 1, we propose a novel neural network-based word embedding model to model the semantic weights from sub-word units. We draw a comparison between Chinese and Indo-European languages in how the semantic weights of sub-words change over time, and show that the weights of characters in Chinese (字 *zi*, the basic sub-word unit in Chinese) are higher in ancient Chinese and lower in modern Chinese, while the opposite trend is observed in Indo-European languages. This is in accordance with theories about monosyllabic-to-bisyllabic shift in Chinese, and the synthetic-to-analytic shift conjecture in Indo-European languages. In Study 2, we apply a different embedding model on another corpus to confirm the finding in Study 1. Although the chronological pattern of semantic weight found is inconsistent with that in Study 1, the results are still meaningful in having discovered the presence of historical changes of sub-word level semantic weights across different corpora and languages.

Our chapter calls for more systematic studies of the applicability of computational embedding methods in modeling the sub-word semantics. Although discrepancies are found in current models and corpora, our empirical findings suggest that word level semantic composition is a dynamic process which reflects historical changes.

Yang Xu & Zheng-sheng Zhang. 2021. Historical changes in semantic weights of sub-word units. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 169–187. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040310

1 Introduction

The roles that sub-word units play in determining word semantics differ across languages. In typical alphabetic languages, such as English, the smallest grammatical sub-word unit is *morpheme* (Katamba 2015). A morpheme is either free or bound: the former stands by itself as a word (e.g., the *root* of English words), while the latter functions only as part of a word (e.g., *affixes* such as *-ness*, *un-*, etc.). In East-Asian languages, however, the distinction between morphemes and words is not as clear. Particularly in Chinese, the basic sub-word unit that acts as a morpheme is the character (字 *zi*), but whether a single morpheme or a combination of morphemes constitute a word is open to debate (Hsieh 2016).

In this chapter, we present two studies that use sub-word incorporated word embeddings to explore the temporal patterns of the semantic weight of sub-word units. In Section 3, we present our first study, in which a novel DYNAMIC SUB-WORD-INCORPORATED EMBEDDING (DSE) model is proposed, which quantifies the semantic weights of sub-word units automatically via joint training tasks. The advantage of this method is that the weights for different words are modeled separately, which provides more fine-grained information. In Section 4, we present the second study, in which we examined the existing model CHARACTER-ENHANCED WORD EMBEDDING (CWE) to obtain sub-word embeddings, and then computed the semantic weights by comparing the norms of sub-word vectors with word vectors. This method leads to faster training and more interpretable results. The purpose of the second study is to confirm whether consistent findings can be reached with a different model and corpus. With these two studies, our goal is to reach reliable conclusions with computational approaches about how the semantic weights of sub-word units change historically.

2 Related work

2.1 Learning vector representations of words

Among the massive amount of approaches to learning dense word vectors, one of the most popular methods is the word2vec model, which implements two efficient ways of learning word vectors, skipgram and CBOW (continuous bag of words) (Mikolov, Sutskever, et al. 2013, Mikolov, Chen, et al. 2013). Both models learn word embeddings by training a network to predict words that co-occur within a window. CBOW aims at predicting the target word given context words in a fixed window, while skipgram predicts the context word given the target word at the center, by maximizing the probability of target/context word, which is approximated with hierarchical softmax or negative sampling (Mikolov, Sutskever, et al. 2013, Mikolov, Chen, et al. 2013).

2.2 Word embeddings with sub-word information

For most languages in the world, the internal structure of words contains information about the semantics of the word. Incorporating parameters associated with those internal structures in the training process can improve word embeddings so that they are more expressive of the meanings of words. There are two types of improvement, semantic compositionality and reducing sparsity. Some languages have strong compositionality at the word level. In Chinese for example, the meaning of a word can be inferred by assembling the meanings of all characters. For instance, the word 教育 *jiao yu* ‘education’, can be inferred from the meanings of its first character 教 *jiao* ‘teach’ and second character 育 *yu* ‘raise’. Based on this, Chen et al. (2015) propose a character-enhanced word embedding model (CWE)

The second type of improvement uses the fact that in some morphologically rich languages, one word can have multiple forms that occur rarely, making it difficult to learn good representations for them. For example, Finnish has 15 cases for nouns,[1] while French or Spanish have more than 40 different inflected forms for most verbs. A way to deal with this sparsity issue is to use sub-word information. Bojanowski et al. (2017) propose to learn representations for character *n*-grams and represent words as the sum of their *n*-gram vectors.[2] Their model, *fastText*, alters the training objective of skipgram by replacing the target word vector with the sum of its *n*-gram vectors.

2.3 Word embeddings and language change

Word vectors have been used to study the long-term change of languages from multiple angles. The most straightforward method is to group text data into time bins and then train embeddings separately on these bins (Kim et al. 2014, Kulkarni et al. 2015, Hamilton et al. 2016). Conclusions about language change are reached by observing how the vectors of the same words change over time. The problem with this approach is that the learned word vectors are subject to random noise

[1] See http://jkorpela.fi/finnish-cases.html.

[2] Another approach is to tokenize words into sub-words while optimizing a language model acquired over these *word pieces* (Schuster & Nakajima 2012, Sennrich et al. 2015).

due to corpus size. Bamler & Mandt (2017) address this with a probabilistic variation of word2vec model, in which words are represented by latent trajectories in the vector space, and the semantic shift of words is described by a latent diffusion process through time. Most of the existing approaches describe language change by the trajectories of some representations in a high dimensional space. Even though this provides rich information about every single point in the space (word, character etc.), it is difficult to interpret and summarize these models and discover the general patterns of language change. Other studies using word embeddings or related methods have been used in very similar context (Tahmasebi et al. 2018, Kutuzov et al. 2018). This chapter explores the historical changes of sub-word level semantics, which has not been studied extensively in existing computational studies.

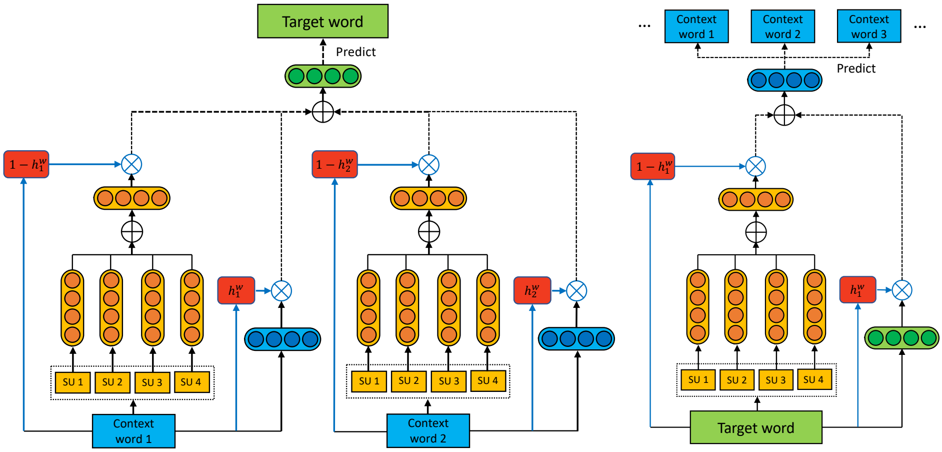
3 Study 1: Relationship between semantic weight and word age

3.1 Dynamic sub-word-incorporated embedding model (DSE)

We propose the dynamic sub-word-incorporated embedding (DSE) model, which captures the semantic weights carried by the sub-word units in words, on top of the architecture of CWE and fastText models. The “dynamic” part is reflected in the design considering that words rely on their internal structures to different degrees in composing a meaning: we associate each word in the vocabulary with a scalar parameter h^w, within the range $[0, 1]$, which is the weight of the word itself in predicting the co-occurring words within a context window. Correspondingly, $1 - h^w$ is the weight of its sub-word units. Here the sub-word units refer to characters in a Chinese word, and a subset of *n*-grams of a word for English and four other languages used in this study. We did not use word roots and affixes as the sub-word units as in Xu et al. (2018), because of the lack of dictionary data in some languages, and the relative simplicity of *n*-gram-based models.

In DSE, we use h^w to compute the weighted average vector for each word, and substitute it for the average context vector x_k in CWE (eq. 5.2), and for the average target vector, as shown below:

$$\begin{cases} x_k' = h_k^w v_k + (1 - h_k^w)\Big(\frac{1}{N_k}\sum_{t=1}^{N_k} c_t\Big), \\ \qquad \text{replacing the } x_k \text{ in eq. (5.2)} \\ x_i' = h_i^w v_i + (1 - h_i^w)\sum_{t=1}^{N_i} c_t, \end{cases} \tag{5.1}$$



(a) DSE-CBOW

(b) DSE-SG

Figure 5.1: The architecture of the two versions of the DSE model. DSE-CBOW associates a semantic weight parameter h^w to each context word, and DSE-SG does this to each target word. The “SU”s in the yellow box stand for “sub-word units”.

in which the subscripts k and i are the indices of words in the vocabulary. We have two versions of model architectures: one is based on CWE (CBOW-like), and the other is based on fastText (skipgram-like). They are referred to as *DSE-CBOW* and *DSE-SG* respectively. The architectures of these models are shown in Figure 5.1.

We call h^w the semantic weight parameter. It describes the proportion of contribution from each word as an unanalyzable semantic unit, while $1-h^w$ is the total contribution from all the sub-word units. h^w is a learnable parameter in the model.

3.2 Corpus data and training setup

We use the Wikimedia database dumps[3] (up until July 2017) as our training data. Data in six languages are used: Chinese (ZH), English (EN), French (FR), German (DE), Italian (IT) and Spanish (ES). Raw text data are extracted from the dump files using `WikiExtractor`.[4] Further text cleaning is conducted by separating sentences into lines, and converting non-proper-nouns (proper-nouns are

[3] https://dumps.wikimedia.org/

[4] https://github.com/attardi/wikiextractor

identified using a pre-trained NER model provided in the Python package `spacy`[5]) to lower case. For Chinese data particularly, word segmentation is carried out using the `Jieba` segmenter.[6] All traditional Chinese characters are converted to simplified characters using OpenCC[7]. All non-Chinese characters are removed, keeping only those within the Unicode range U+4E00–U+9FFF. The training data of all six languages are of similar volumes: 33 to 40 million tokens each after preprocessing.

To accelerate training, we limit the number of effective semantic units in each word. For Chinese data, words containing more than 7 characters are ignored. For other languages, if a word contains more than 7 *n*-grams, we randomly select 7 out of them, and ignore the rest. Here the number 7 is chosen based on the following empirical observation: in a pilot study, we found that numbers larger than 7 will not improve the resulting embeddings, but significantly slow down the training. Other hyper-parameters are kept as close to the previous studies as possible. The values of the hyper-parameters for training the DSE models are shown in Table 5.1.

Table 5.1: Hyperparameter setting for Study 1.

Hyperparameter	Value
Embedding size, word	300
Embedding size, sub-word	300
Window size	5
Number of negative samples	10
Batch size	128
Minimal word frequency	5
Initial learning rate, DSE-CBOW	0.05
Initial learning rate, DSE-SG	0.025

The training stage consists of three steps:

1. Pre-train the word embeddings: set the parameters for word embeddings, i.e., the v_k and v_i in Equation (5.1) to trainable; set all the other parameters to not trainable; train the model for 5 epochs.

[5] https://spacy.io/

[6] https://github.com/fxsjy/jieba

[7] https://github.com/BYVoid/OpenCC

2. Pre-train the sub-word embeddings: set the parameters for sub-word units, i.e., c_t in Equation (5.1) to trainable; set all the other parameters to not trainable; train the model for 5 epochs.

3. Set all the parameters to trainable (including embeddings and h^ws); train the model for 5 epochs.

As for the size of n-grams, we use a fixed size $n = 4$, i.e., no bigrams or trigrams are considered. This choice is partially based on Bojanowski et al.'s (2017) work showing that $n = 4$ already achieves a satisfactory embeddings, and partially due to speed consideration. For words that consist of more than 4 letters, we only consider two sources for the mixture embeddings: the word itself and the n-gram ($n < 4$).

The semantic weight parameters h^w are implemented as a $V_w \times 1$ lookup table. Thus, in each training step, the learning algorithm updates three embedding tables: word embeddings E_w, character embeddings E_c, and the semantic weights. Specifically, for the DSE-SG model, the average embeddings are first computed from E_w, E_c, h^w, and h^c using eq. (5.1) and then outputted as the final word vectors. For DSE-CBOW model, just the E_w table is outputted as the learned word vectors.[8]

3.3 Results and discussion

We are interested in examining the relationship between the semantic weight h^w of a word and its relative "age". According to the observation that Chinese is shifting from monosyllabic words to bisyllabic words, it is reasonable to expect that newer Chinese words should have larger h^w than older words, because a higher h^w indicates that the word as a whole rather than the individual sub-word units is more important in determining its meaning. For other languages, we do not have a clear idea on what the relationship could be, but they should provide an interesting comparison.

First, we need to have a reliable way to measure the "age" of a word. We use the Google Books Ngram (GBN)[9] corpus, which contains word frequency information from about 10 million books published over a period of five centuries (Lin et al. 2012). It is the best resource we can find that provides estimated temporal distributions of words in multiple languages. For each word in GBN we

[8] The discrepancy exists in the original implementations of CWE and fastText, and the reason for it is out of the scope of this study.

[9] http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

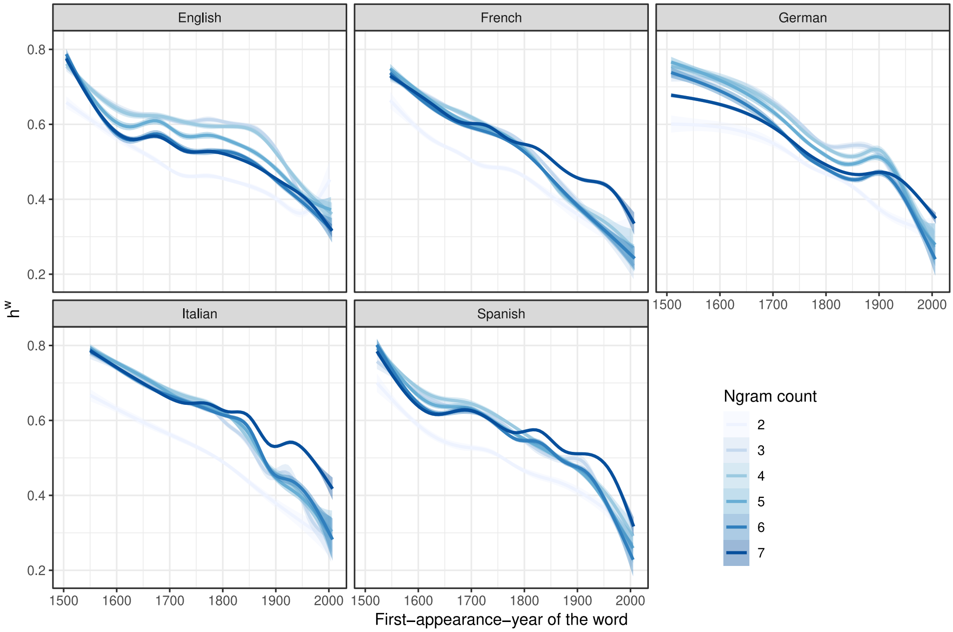


Figure 5.2: Semantic weight h^w against the first-appearance-year of words in DE, EN, ES, FR, and IT. Words with sub-word units (n-grams) number ranging from 2 to 7 are plotted separately. Shaded area indicates 95% point-wise confidence intervals of the fitted regression lines. h^w scores are from the DSE-SG model.

extract the first *year* it appears in the dataset, and use this first-appearance-year as an approximation of the word's age. Then we check if the word's age is correlated with its h^w from training the DSE model. For example, the word 爱人 *ai ren* 'lover' first appears in 1804 CE (at least according to the GBN collection). Thus, our examination is based on the intersection of vocabularies between GBN and the training data. For DE, EN, ES, FR and IT, the intersection covers above 95% of the most common words in the training set, and the proportion for ZH is 84%.

In a short summary of the results, we find opposite $h^2 \sim$ year relationships in Chinese and the other five languages. h^w decreases with the first-appearance-year in the five Indo-European languages, as shown in Figure 5.2. Words with sub-word units count ranging from 2 to 7 are included. Short words that have only 1 n-gram are excluded because the n-grams have the same form as the words. There are some fluctuations but the overall decreasing trends of h^w are salient. As the decrease of h^w is equivalent to the increase of $1 - h^w$, it indicates that in these five languages, sub-word units carry more semantic weights in newer words than older ones. The h^w scores reported in Figure 5.2 are from the DSE-SG model.

As for Chinese, however, h^w increases with the first-appearance-year as shown in Figure 5.3. We choose the sub-word units (characters) count = {2, 3, 4} because they are the majority in the training data, with proportions 57.5%, 31.0%, and 8.6%. Frequency-wise, their proportions are more dominant: 82.9%, 11.8%, and 4.6% respectively. Single-character words are excluded because the vast majority (98%) of words in the training data are multi-character ones. Words composed of more than 4 characters are very uncommon in Chinese. From the plot, the increasing trends of the 2-character words are observable, but less so for the 3- and 4-character words. This indicates that our hypothesis is supported: characters carry more semantic weight in older Chinese words than in newer Chinese words.

Besides, an interesting finding is that the h^ws from DSE-SG are larger than those from DSE-CBOW in Chinese. It makes sense intuitively: a CBOW-like model is using multiple context words to predict one word, and thus the semantic weight of each individual word is diluted.

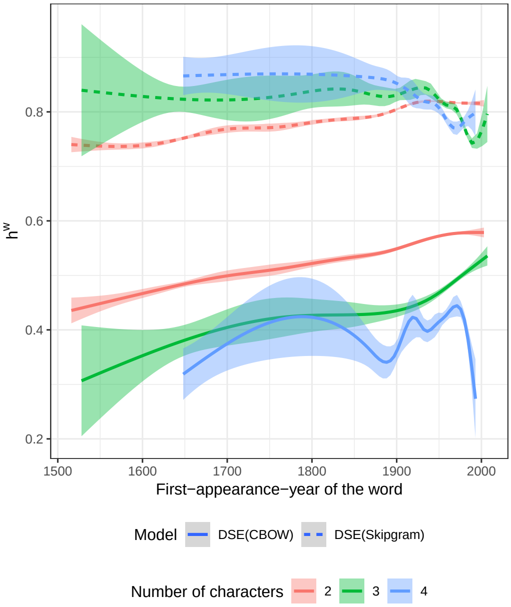


Figure 5.3: Semantic weight h^w against first-appearance-year for Chinese words with character number = 2, 3, and 4. Shaded area indicates 95% point-wise confidence intervals of the fitted regression lines.

4 Study 2: Temporal patterns of semantic weight in historical corpora

In this study, we collect text data from Wikisource, a public resource of historical articles. We divide the text data into segments according to the years of authorship, and train embedding models on each segment individually. These individual models can reflect the semantic weight for each historical period, and we carry out a longitudinal analysis on how the semantic weight evolves.

4.1 Character-enhanced word embedding (CWE) model

The model we utilize in this study is the character-incorporated word embedding models (CWE) (Chen et al. 2015), which presents modifications on top of the original word2vec model. The design goal of CWE is to obtain a richer representation of Chinese words by assigning a vector to each character in a word. It replaces the context word vector, with an average vector x_k,

$$x_k = \frac{1}{2} v_k + \frac{1}{2}\Big(\frac{1}{N_k}\sum_{t=1}^{N_k} c_t\Big) \tag{5.2}$$

where N_k is the number of characters in word w_k, and c_t is the vector of the tth character. Here the weights on the word and the characters within that word are equal (0.5), which is based on an empirical hypothesis that context words and characters are equally important in determining the semantics of target word.

4.2 Data collection and preprocessing

Wikisource[10] is part of the Wikimedia foundation,[11] which has the stated goal of developing and maintaining open content, wiki-based projects and providing the full contents of those projects to the public free of charge. It hosts text data from a broad range of categories and timespans, including professionally published articles, newspaper articles, archived documents, etc. Wikisource includes multiple language-specific sub-domains with each article labeled with “author”, “title”, and “publication time” (with a yearly granularity). The largest sub-domains in terms of article number are English, French, Chinese, German, Spanish, and Russian. Thus, we include these six languages in this study. The ProofreadPage extension makes sure that all the works on the website are verifiable, reliable, and

[10] https://wikisource.org

[11] https://wikimediafoundation.org/

accurate. Wikisource provides ancient (600 CE) as well as contemporary articles. We therefore consider Wikisource a useful resource for building a corpus for historical language studies. Regrettably, only a few researchers have conducted research using this material.

Wikisource does not offer direct download links of the data, so one of the challenges of this project is to acquire the textual data from the website. Furthermore, anyone can edit articles on the website, so the structure of each HTML page differs from the others. In order to solve this irregularity issue, distinctive web crawlers for each subdomain were developed and the crawled JSON data was extracted into text documents.

The collected corpus contains articles from the 11th century to the 21st century; however, the number of articles is not evenly distributed along the timeline. The amount of textual data for each bin is very important for providing an accurate description of the semantics of a language for that time period. To overcome this difficulty, we will only consider the articles dated from 1820 to 1930. These articles were divided into temporal bins of 10 years. This division is arbitrary and it does not correlate with any semantic difference in the language. For this study, we use the Chinese subset of the corpus, because the target model CWE is designed for Chinese language only.

4.3 Word segmentation

The Chinese written language is printed without marking boundaries between words, like the blank space that is commonly used in other languages. Thus, it requires a preprocessing step known as word segmentation, which places boundaries between adjacent characters in order to identify the unit of "word". We use the `jieba` word segmentor[12] for this study. Although `jieba` is not designed for ancient Chinese, we found that it is able to detect words that belong to ancient vocabulary, such as 中书 *zhong shu* 'an official position during the Tang dynasty', 若夫 *ruo ru* 'if' etc. The resulting corpus data with various word counts, character counts, and vocabulary sizes in terms of unique word tokens can be seen in Figure 5.4.

4.4 Definitions of semantic weight

We define *semantic weight* in a different way from that of Study 1 (Section 3). Here, it is defined as the proportion of the Euclidean norm of a word vector relative

[12] https://github.com/fxsjy/jieba

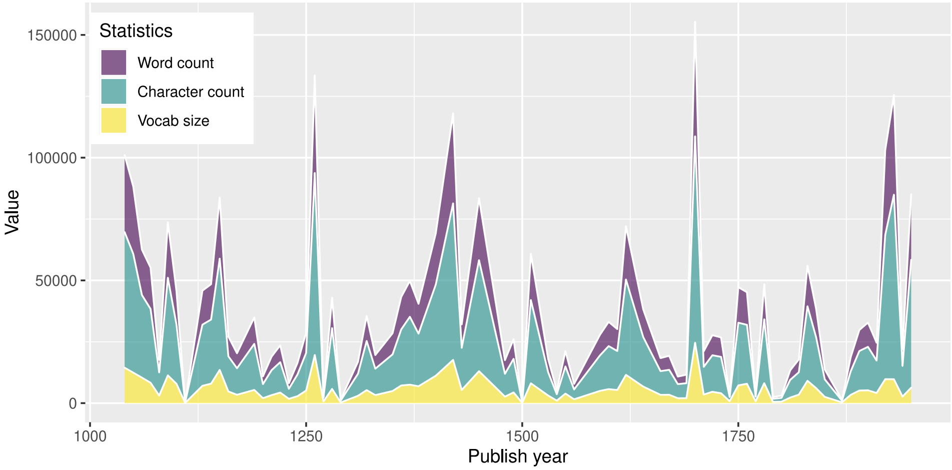


Figure 5.4: The distribution of word count, character count, and vocabulary size (number of unique words) across article publication years in the Wikisource corpora.

to the mean norms of its constituent character vectors, designated by the Ω in Equation (5.3),

$$\Omega_w = \frac{\|v_w\|}{\frac{1}{N_w}\sum_{c\in w}\|v_c\| + \|v_w\|} \tag{5.3}$$

where N_w is the number of characters in word w. v_c is the embedding vector of character c, and v_w is the word embedding vector. $\|v_w\|$ and $\|v_c\|$ are the Euclidean norms, which have theoretically unbounded positive values. A word with larger word vector norm $\|v_w\|$ will have a larger Ω_w score, while a word with a larger mean character norm $\|v_c\|$ will have a smaller Ω_w score. Thus, Ω_w quantifies the degree to which a word functions as a whole semantic unit as opposed to its constituent sub-units. Since CWE has two versions of implementation, CBOW and Skipgram based, we examine both and use CWE-CBOW and CWE-Skipgram to refer to the models respectively.

4.5 Model training procedure

We first split the training data into segments, based on the publication year of the individual articles, and train one embedding model for each segment. We need to choose the size of segments carefully, because we need to have sufficient number of segments in order to find a consistent temporal pattern, while it is also necessary to make sure that each segment is sufficiently large so that the embedding models are effectively trained. It is basically a trade-off between granularity and effectiveness.

The whole training set, designated by $\mathcal{D}$, is segmented into N segments, resulting in $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$. We experimented with $N = 5$ and $N = 10$. Since the token numbers are not evenly distributed among years, the size of $\mathcal{D}_i$ varies. In order to eliminate the potential confounding effects due to the varying sizes of training data, we randomly sample 30k lines of text from each $\mathcal{D}_i$ into $\mathcal{D}'_i$.

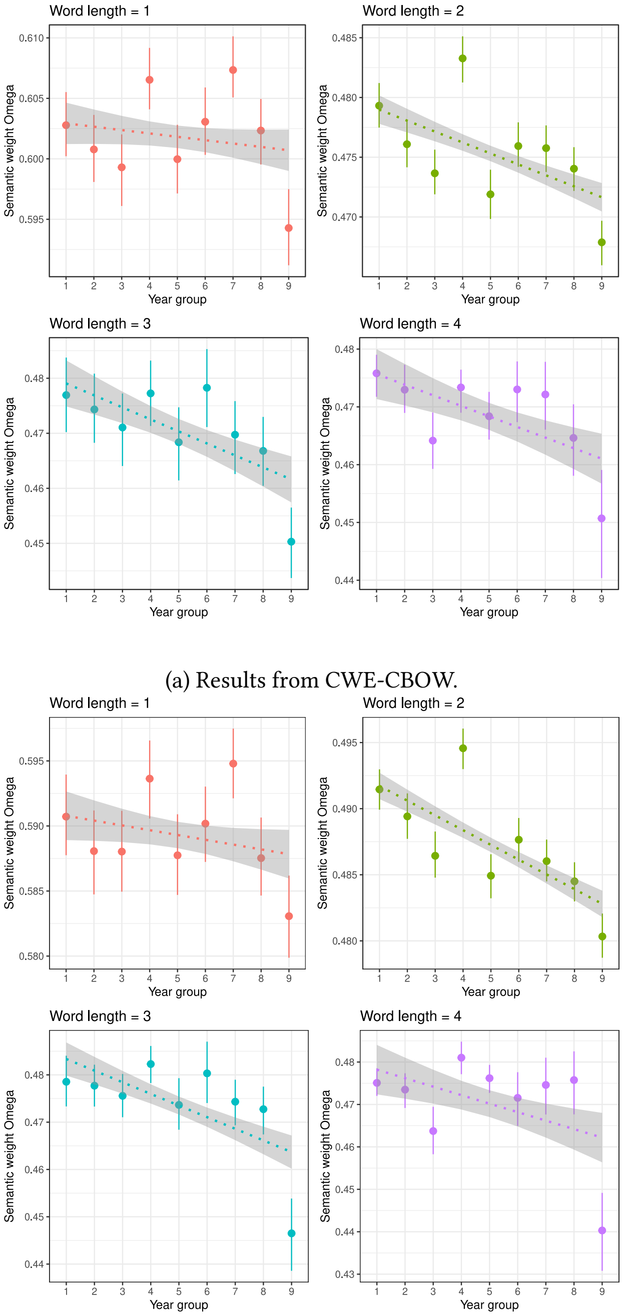
For each $\mathcal{D}'_i$ in $\mathcal{D}$ ($i = 1, \ldots, N$; $N = 9$), we train a CWE model, and calculate the Ω_w score for each word in vocabulary V'_i. Then we use the mean score $\Omega_i = \frac{1}{T_i} \sum_{w \in V_i} \Omega_w$ to estimate the average semantic weight of historical period i. The purpose is to examine the relationship between Ω_i and i. Our assumption is that a correlation between Ω_i and i should be observed.

4.6 Results and discussion

We plot the Ω_i score against the historical period i in Figure 5.5. It can be seen that Ω_i decreases as i increases. Because larger values of i represent the historical periods closer to modern time, the observed increasing trend indicates that words in modern languages have smaller Ω_i than words in ancient times. The same trends hold for both CWE-CBOW (Section 4.6) and CWE-Skipgram (Section 4.6).

In order to make a less biased comparison, in Figure 5.5 we individually observe words of different lengths, $l = 1, 2, 3, 4$. The $l = 1$ group includes mono-character words, such as 一 *yi* ‘one’, 三 *san* ‘three’, 万 *wan* ‘ten thousand’, etc. The $l = 2$ group includes bi-character words, such as 一定 *yi ding* ‘must’, 不能 *bu neng* ‘cannot’, 世事 *shi shi* ‘world affairs’ etc. The $l = 3$ and $l = 4$ groups contain more proper nouns (person and organization names), and fixed idioms, such as 士大夫 *shi da fu* ‘scholars’, 皇太后 *huang tai hou* ‘queen’, 都督府 *du du fu* ‘governor’s office’, etc. We fit individual linear models with formula $\Omega_i \sim i$ for all four length groups, and all models return statistically significant negative coefficients ($p < 0.05$), indicating that the observed decreasing trends are reliable. The fitted regression lines and the bootstrapped 95% confidence intervals are shown in Figure 5.5.

Another fact worthy of note is that only a small subset of words appears throughout the whole time span. For the accuracy of demonstration, we include only those words that exist in all nine vocabulary sets $V_i (i = 1, 2, \ldots, 9)$, which we refer to as *common vocabulary*. The size of the common vocabulary is negatively correlated with word length, which is expected as an prediction of Zipf’s law (Li 1992).



(a) Results from CWE-CBOW.

(b) Results from CWE-Skipgram.

Figure 5.5: Average semantic weight Ω_i in nine (9) historical groups ($i = 1, 2, \ldots, 9$). The results from CWE-CBOW and CWE-Skipgram are shown in (a) and (b) respectively.

It is surprising to find that the Ω metric demonstrates contrary patterns as compared with the h^w metric in Study 1. The semantic weight Ω demonstrates an opposite pattern compared with the h^w coefficient defined in Study 1. In Study 1, the main finding about Chinese described in Figure 5.3 shows that the semantic weight measured by h^w increases with the age of a word. The Ω in Study 2 shows a clear decreasing trend with historical period.

However, we do not think the results from Study 2 are sufficient to totally reject the conclusion from Study 1. First, the x axis in Study 1 is the approximated "age" of a word, acquired from an external dictionary book (Google Books Ngram), while the x axis in Study 2 is the *actual* publication year. The independent variables of the two studies are essentially different. Based on the results, we lean towards Study 2 because the decreasing pattern of Ω in Chinese is consistent with those of the five Indo-European languages (Figure 5.2). We suspect that the age of Chinese words according to GBN may not be an accurate estimate. Secondly, the way we obtain h^w and Ω_w is different as well. h^w is automatically learned from data during the training stage of DSE model, while Ω_w is a post-hoc quantity computed after the CWE model is trained. In theories of representation learning (Bengio et al. 2013), more informative parameters are assigned with larger weights by the model, thus h^w and Ω_w should bear the same semantic weights. Based on these considerations, we conjecture that the discrepancy between Studies 1 and 2 is primarily due to the different operational definitions of historical periods. Beyond that, the empirical findings from both studies clearly indicate that the semantic weights of sub-word units indeed change with historical periods, confirmed by multiple corpora and models.

5 General discussion and conclusions

The findings from Study 1 provide new evidence to linguistic theories about word formation. First, what constitutes a word in Chinese has changed: compared to its earlier stage, modern Chinese tends to have multiple characters for a single semantic unit. The semantic weight carried by a single character is decreasing. This is strong evidence in favor of the claim in qualitative studies that Chinese has been evolving towards multisyllabicity from monosyllabicity. Second, the trend of increasing semantic weights on sub-word units in Indo-European languages is consistent with the "synthetic → analytic" pattern shift at the phrase level composition (Hamilton et al. 2016). Moreover, the relative "synthetic" way of composing Chinese word found in this study seems consistent with the holistic encoding hypothesis in the perceptual theories about the Chinese writing system (Dehaene et al. 2005, Mo et al. 2015).

However, the above conclusions are not directly supported by the findings from Study 2. Both the Ω_w and h^w quantify the role that a word itself as an atomic unit is playing in contributing to the semantic meanings, when sub-word units are also contributing to the meaning. Ω and h^w should be of smaller value if sub-word units carry critical semantic information; they should be of greater value if sub-word units are not contributing actively. Thus, we believe the magnitudes of both quantities should correctly reflect the semantic importance played by sub-word units. Purely from the results of Study 2, we can also argue that the individual characters in Chinese are playing more and more important roles as the language evolves. The inconsistency between Study 1 and 2 is primarily due to the different ways of setting up historical periods. In Study 1, we use the first year in which a word appears in a large collection of printed materials, which is less accurate than the segmentation method by actual publication year in Study 2.

The usage of Google Books Ngram (GBN) dataset in Study 1 can be the direct cause for the inconsistency from Study 2. The lexicon publication year information in GBN is obtained from the OCR scans, which may suffer from missing pages or misrecognition. The main advantage of GBN is its support for multiple languages. For future work, more accurate resources for identifying word ages should be explored. For example, the Oxford English Dictionary (Simpson & Weiner 1989) is a better resource for English, as it records the ambiguities and semantic changes for a large vocabulary of English, which can be used to identify the “birth” year of specific word meanings. Another planned improvement is to extend the range of sub-word units explored other than morphemes, for example, semantically-associated sub-word units such as phonesthemes (Bergen 2004, Sagi 2019), sound symbolism (Imai et al. 2008) etc., because we assume the sub-word level semantic decomposition is ubiquitous, and should go beyond the predefined concepts of morphemes.

Regardless of the seemingly conflicting results of the two studies presented in this chapter, we believe some meaningful empirical findings are discovered. First of all, the semantic weight of sub-word units can be quantified by well designed computational models. The parameters in those unsupervised machine learning models can provide interesting information that is not available with other count based statistical tools. Though we need to be careful when choosing proper models and proper ways of defining the computational metrics of semantic weights in future studies. At least fine-grained embedding models such as DSE and CWE should be further examined in terms of their behavioral consistency. More importantly, the semantic weights of sub-word units indeed demonstrate a clear pattern of change along historical periods, which to the best of our knowledge, is not discussed in previous studies. The semantic weights defined in this study can be viewed as a metric of the "atomicness" of words. We put forward a dynamic theory of word and sub-word level semantic composition – the way we compose words, invent new words, and reuse old words, can be governed by some universal rules. What these rules are, and how they are related to the linguistic capacity of human beings are the research questions that await future work.

Acknowledgements

We thank Nika Nizharadze for his effort in collecting and preprocessing the text data. We are grateful to Adam Jatowt for his great help in the preparation of this chapter. We appreciate the hard work from all the reviewers and editors.

Abbreviations

CBOW	continuous bag-of-words
CWE	character-enhanced word embedding
DSE	dynamic sub-word-incorporated embedding
GBN	Google Books Ngram
HTML	Hypertext Markup Language
JSON	JavaScript Object Notation
NER	named entity recognition
OCR	optical character recognition
SG	skipgram

References

Bamler, Robert & Stephan Mandt. 2017. Dynamic word embeddings. In Doina Precup & Yee Whye Teh (eds.), *Proceedings of the 34th international conference on machine learning* (Proceedings of Machine Learning Research 70), 380–389. Sydney: PMLR.

Bengio, Yoshua, Aaron Courville & Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8). 1798–1828.

Bergen, Benjamin K. 2004. The psychological reality of phonaesthemes. *Language* 80(2). 290–311.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL* 5. 135–146.

Chen, Xinxiong, Lei Xu, Zhiyuan Liu, Maosong Sun & Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *Proceedings of IJCAI 2015*, 1236–1242.

Dehaene, Stanislas, Laurent Cohen, Mariano Sigman & Fabien Vinckier. 2005. The neural code for written words: A proposal. *Trends in Cognitive Sciences* 9(7). 335–341.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Hsieh, Shu-Kai. 2016. Chinese linguistics: Semantics. In Sin-Wai Chan, James W. Minett & Florence Li Wing Yee (eds.), *The Routledge encyclopedia of the Chinese language*, 203–214. Abingdon: Routledge.

Imai, Mutsumi, Sotaro Kita, Miho Nagumo & Hiroyuki Okada. 2008. Sound symbolism facilitates early verb learning. *Cognition* 109(1). 54–65.

Katamba, Francis. 2015. *English words: Structure, history, usage*. London: Routledge.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore: ACL. DOI: 10.3115/v1/W14-2517.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*, 1384–1397. Santa Fe: ACL.

Li, Wentian. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory* 38(6). 1842–1845.

Lin, Yuri, Jean-Baptiste Michel, Erez Aiden Lieberman, Jon Orwant, Will Brockman & Slav Petrov. 2012. Syntactic annotations for the Google Books NGram corpus. In *Proceedings of the ACL 2012 System Demonstrations*, 169–174. Jeju Island: ACL. https://www.aclweb.org/anthology/P12-3029.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint 1301.3781.*

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in neural information processing systems*, 3111–3119. Red Hook, NY: Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mo, Ce, Mengxia Yu, Carol Seger & Lei Mo. 2015. Holistic neural coding of Chinese character forms in bilateral ventral visual system. *Brain and Language* 141. 28–34.

Sagi, Eyal. 2019. Taming big data: Applying the experimental method to naturalistic data sets. *Behavior Research Methods* 51(4). 1619–1635.

Schuster, Mike & Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *Proceedings of ICASSP 2012*, 5149–5152. IEEE.

Sennrich, Rico, Barry Haddow & Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909.*

Simpson, John & Edmund Weiner (eds.). 1989. *The Oxford English Dictionary.* 2nd edn. http://dictionary.oed.com.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278.*

Xu, Yang, Jiawei Liu, Wei Yang & Liusheng Huang. 2018. Incorporating latent meanings of morphological compositions to enhance word embeddings. In *Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, vol. 1, 1232–1242.

Chapter 6

Chaining algorithms and historical adjective extension

Karan Grewal & Yang Xu

University of Toronto

Natural language relies on a finite lexicon to express a potentially infinite set of ideas. This tension often results in the innovative reuse of existing words to describe emerging ideas. In this chapter, we take a computational perspective to examine how English adjectives extend their range over time to modify nouns and form previously unattested adjective-noun pairs. We hypothesize that how novel adjective-noun pairings emerge is non-arbitrary and follows a process of chaining, whereby novel noun referents for an adjective link to existing nouns modified by the same adjective that are close in semantic space. We test this proposal by exploring a set of probabilistic models that predict adjective-noun pairs from a historical text corpus (Google Books) that spans the past 150 years. Our findings across three diverse sets of adjectives support a chaining mechanism sensitive to local semantic neighbourhood – formulated as an exemplar model of categorization similar to the Generalized Context Model. These findings mirror existing work on chaining in the historical growth of grammatical categories. We discuss the limitations and implications of our approach toward a general theory of word meaning extension in natural language.

1 Introduction

Natural language relies on a finite lexicon to express a potentially infinite set of ideas. One result of this tension is the innovative reuse of existing words (Ramiro et al. 2018). Here we explore how English adjectives extend their range over time to modify novel nouns and ask whether there are principled mechanisms in the historical process of adjective extension.[1]

[1] See Grewal & Xu (2020) for a shorter conference version of this work.

Karan Grewal & Yang Xu. 2021. Chaining algorithms and historical adjective extension. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 189–218. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040312

The topic of adjective-noun composition has been discussed in the computational literature. Existing studies have explored which adjective-noun pairings are considered plausible (Lapata et al. 1999), and how adjectives can be combined with nouns sensibly either via probabilistic models (Lapata 2001) or through ontological constraints (Schmidt et al. 2006). Recent work has also suggested that adjective-noun composition can be modelled using vector-space models such as Word2Vec (Mikolov et al. 2013). In these studies, adjectives are considered to be linear operators that act on nouns in a vector space that impose linear transformations (Baroni & Zamparelli 2010, Boleda et al. 2013, Vecchi et al. 2013, 2017) or conform to additive compositional models (Zanzotto et al. 2010). Despite this extensive line of work, sparse computational research has considered the dimension of time in the investigation of adjective-noun composition.

Independent research in historical linguistics has explored adjective extension from the perspective of semantic change. In particular, Williams (1976) studied meaning change in synaesthetic adjectives and found that sensory terms such as those pertaining to sound, touch, and smell exhibit regular semantic shift such that words from the same sensory domain tend to undergo parallel change in meaning. For instance, Williams (1976) showed how adjectives that originally described the sense of touch have since extended to describe color (e.g., *warm cup → warm color*), and adjectives that originally described color have later extended to describe ideas associated with sound (e.g., *clear blue → clear voice*). This line of inquiry takes an empirical approach to characterize meaning change in adjectives from a focused semantic domain, but to our knowledge the more general problem of how adjectives extend their range to describe novel noun referents has not been treated formally or explored at scale.

We investigate whether adjective extension might follow non-random processes that make novel adjective-noun pairings yet to emerge in a linguistic community predictable. Our view is that novel adjective-noun pairings provide an incremental way of extending the referential range of adjectives, and word meaning extension or semantic change might result from this process (e.g., consider meaning extension in the adjective *cold* reflected in a chain of different noun context: *cold food → cold person → cold war*). It is conceivable that pairing with novel nouns does not necessarily entail semantic change in an adjective (e.g., *cold Gatorade* does not entail semantic change in *cold* which had the meaning ‘low-temperature’, even though *Gatorade* might appear as a novel item to pair with *cold* at some point in history), and our main focus here is to characterize the general mechanisms of an adjective’s extension over time – with or without semantic change.

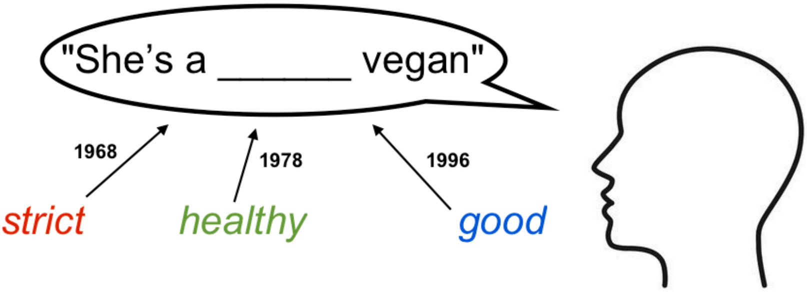


Figure 6.1: Example adjectives that emerged to describe *vegan* over the past half century.

Figure 6.1 illustrates that historical adjective-noun pairings can often be subject to non-linguistic or external influences which make them non-trivial to predict. For instance, the emergence of *vegan* is largely a cultural product, and different adjectives have been extended to modify this noun over time presumably as a result of cultural development. Our premise is that despite the historical adjective-noun pairings may be subject to socio-cultural influences, language users must somehow choose adjectives sensibly to describe nouns so that the novel pairings can be related to the original meaning of the adjectives. For this reason, we expect the historical processes of adjective extension to follow non-arbitrary paths.

We formulate adjective extension as a temporal prediction problem: Given adjective-noun pairings at historical time t, can we predict novel adjective-noun pairings into the future at $t + \Delta$? We ground our work in cognitive linguistic theories of chaining, which have been proposed and recently demonstrated as important cognitive mechanisms for historical word meaning extension (Lakoff 1987, Malt et al. 1999, Bybee et al. 1994, Sloman et al. 2001, Xu et al. 2016, Ramiro et al. 2018, Habibi et al. 2020). A consistent finding from these studies is that chaining as an extensional mechanism depends on semantic neighbourhood density, highlighting the fact that historical word meaning extension tends to follow incremental as opposed to abrupt processes. In our study, we consider each adjective as a linguistic category and explore different mechanisms of chaining to predict how adjective categories grow to modify nouns that they have not previously been paired with. We next describe the theory of chaining and related work on word meaning extension.

2 Theory of chaining and word meaning extension

The proposal of semantic chaining is rooted in cognitive linguistic work on categories, or more specifically, radial categories (Lakoff 1987). By this view, chaining is a process of meaning extension whereby novel items link to existing items of a linguistic category due to proximity in semantic space. This process leads to chain-like semantic structures, and Lakoff (1987) has considered it a key mechanism for growing radial categories or semantic networks, i.e., how categories grow "spokes" of meaning from a central core meaning. Lakoff's (1987) original work discusses chaining in a number of exemplary domains such as the grammatical categories of classifiers in Japanese and Dyirbal (an Australian aboriginal language), and prepositions such as how the English spatial term *over* extends over a wide variety of spatial (e.g., *over the hill*) and metaphorical context (e.g., *over the moon*). Later work also discusses chaining in the grammar evolution of tense, modality, and aspect systems (Bybee et al. 1994), container naming (Malt et al. 1999), and metonymical semantic shift (Hilpert 2007). These studies have broadened the view of chaining toward a generic mechanism for grammatical and semantic changes in language, although they do not provide a formal account for the processes of chaining or test this idea comprehensively against historical corpus data.

Extending the cognitive linguistic accounts of chaining, recent work has explored formal approaches to chaining in several aspects. Sloman et al. (2001) and Xu et al. (2016) have developed computational models of chaining and tested the extent to which these models account for the extension of container names such as *bottle* and *jar*. Their findings suggest that chaining depends on semantic neighbourhood density, and more specifically nearest-neighbour models of chaining tend to best account for the empirical data. Ramiro et al. (2018) extend this work to examine whether similar models of chaining might explain the historical emergence of senses (or word sense extension) in English words over the past millennium, e.g., how *face* might extend from 'body part' to senses including 'front surface (of an object)', 'facial expression', and 'defy danger'. Their work confirms the earlier finding that chaining relies on semantic neighbourhood density, and senses tend to emerge by linking those that are close in semantic space.

More recent work has built on these computational studies to investigate the historical growth of grammatical categories, and particularly numeral classifiers commonly used in East Asian languages (Habibi et al. 2020). This work has examined a suite of probabilistic models of chaining and found chaining to be best captured by an exemplar model, also known as the Generalized Context Model in the psychological literature of categorization (Nosofsky 1986). By this view, chaining in linguistic categories reflects an exemplar-based process of extension that mirrors those found in other aspects of language change including phonetics, morphology, word senses, and constructions (Skousen 1989, Pierrehumbert 2001, Keuleers 2008, Bybee 2013, Ramsey 2017).

Here we examine chaining through the lens of the exemplar theory but in a new domain: the case of historical adjective extension in English. Analogous to how numeral classifiers (e.g., in Mandarin Chinese) extend toward novel nouns, English adjectives also extend to modify novel noun referents. If the exemplar view represents a general mechanistic account for the growth of linguistic categories, it should explain the historical extension of adjective categories.

Figure 6.2 illustrates the exemplar theory of chaining with two example adjectives and a dimension-reduced semantic space of their noun referents, data for which were taken from the Google Books corpus (Michel et al. 2011) during the 1880s. The two adjectives *wrong* and *troubled* are closely related in semantic space in the 1880s and share noun referents (labelled in purple) such as *war* and *humanity.* The emergent or query noun *slavery* has not appeared in close context with either adjective prior to the 1880s but is in semantic proximity of their noun referents. The exemplar view of chaining postulates that the linguistic category having a higher local semantic similarity (or neighbourhood density) to a novel referent is more likely to attract that item, and when this process repeats over time chain-like category structures may result in semantic space. Here, *wrong* has a higher neighbourhood density (with its noun referents labelled in red) to *slavery* in comparison to *troubled* (with its noun referents labelled in blue), namely that the existing noun referents of *wrong* are closer in semantic space to the query noun than those of *troubled.* The exemplar view of chaining thus predicts that *wrong* is a more likely adjective candidate to be paired with *slavery*, which aligns with the empirical data. We seek to evaluate the extent to which the exemplar model of chaining accounts for historical adjective extension, and if it is better or worse than alternative accounts for the chaining process.

3 Computational formulation of theory

We formulate adjective extension as a temporal categorization problem and explore the process of chaining via a suite of models that predict adjective-noun pairings over time. The probabilistic formulation we describe here follows existing work on chaining and the extension of numeral classifiers (Habibi et al. 2020).

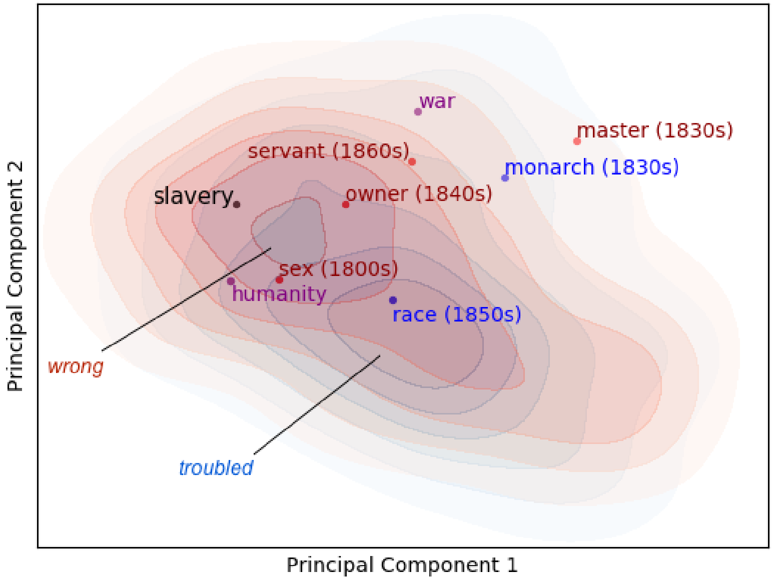


Figure 6.2: An illustration for the exemplar view of semantic chaining (Habibi et al. 2020) using two example adjectives *wrong* and *troubled*. The semantic space is constructed from the first 2 principal components in the Principal Components Analysis on the diachronic Word2Vec embeddings from the 1870s (Hamilton et al. 2016). Nouns labelled in purple (e.g., *humanity*, *war*) are shared context of the two adjectives. Nouns labelled in red (e.g., *master*, *servant*, *owner*, *sex*) and blue (e.g., *monarch*, *race*) are contexts that co-occurred more often with *wrong* and *troubled* respectively up to the 1880s. The contours represent probability distributions of nouns co-occurring with each of the two adjectives, constructed by kernel density estimation.

3.1 Probabilistic formulation

Given an emergent query noun n^* at a future time $t + \Delta$ and a finite set of adjectives $\mathcal{A}$, we seek to predict which adjective(s) $a \in \mathcal{A}$ would be most appropriate for describing n^* at time $t + \Delta$ based on the historically attested adjective-noun pairings at current time t.[2] We cast this problem as probabilistic inference over the space of adjectives for a query noun n^*:

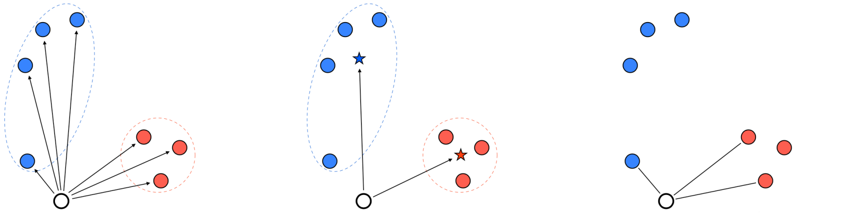
$$p\left(a|n^*\right)^{(t+\Delta)} \propto p\left(n^*|a\right)^{(t)} p\left(a\right)^{(t)}. \tag{6.1}$$

[2] In our formulation of the prediction problem, we consider an adjective-noun pair to be novel if (1) the noun itself is novel or (2) the pairing has not been attested in history.

The posterior term $p\left(a|n^*\right)^{(t+\Delta)}$ relies on two sources of information to predict the choice of adjective(s) for n^*: (1) a likelihood function $p\left(n^*|a\right)^{(t)}$ that specifies the semantic proximity of n^* to an adjective a given knowledge of its existing noun referents at time t, and (2) a prior distribution $p\left(a\right)^{(t)}$ that captures the a priori belief or probability of choosing an adjective a from the current lexicon without considering its semantic relation to n^*. In both our formulations of the likelihood and the prior, we focus on type-based representations of adjective-noun co-occurrence frequencies and adjective frequencies. Token-based representations have been explored and shown to be inferior in accounting for the historical growth of classifier categories in related recent work (Habibi et al. 2020).

3.2 Likelihood function

We describe a suite of models to explore a space of possible candidates for the likelihood function. Each of these models postulates a different mechanism of chaining that links existing noun referents of an adjective to a novel noun that appears at a future time. We use $\{n\}_a^{(t)}$ to denote the semantic embeddings for the set of nouns that co-occur with adjective a at current time t, i.e., the semantic representation for the collective set of noun referents for adjective category a. Figure 6.3 provides an illustration for the representative chaining models that we describe in the following subsections.



(a) exemplar (b) prototype (c) k-nearest neighbours, $k = 3$

Figure 6.3: An illustration of representative chaining models for the likelihood function. The empty circle represents the stimulus or the query noun n^*. Red circles represent nouns that are attested to have paired with one particular adjective, and blue circles represent nouns that are attested to have paired with an alternative adjective (in reality, a noun can pair up with multiple adjectives). The dotted lines indicate the noun referent space for a given adjective. The stars represent the prototypes under the prototype model. The lines indicate the influence of existing (exemplar) nouns to the query noun as specified in each model of chaining.

3.2.1 Exemplar model

The first likelihood function we consider is based on the exemplar theory which is discussed in the psychological literature of categorization (Nosofsky 1986). Here each noun $n \in \{n\}_a^{(t)}$ is treated as an exemplar for an adjective a.

The exemplar view of chaining postulates that a query noun should be linked to an adjective category where the noun exemplars are most proximal in semantic space. As such, a novel noun is pulled or attracted to the adjective category that has the highest local semantic density around that noun. The likelihood term between n^* and adjective a is thus proportional to the weighted sum of similarities between n^* and the noun exemplars of a:

$$p(n^*|a)^{(t)} \propto \frac{1}{h\left|\{n\}_a^{(t)}\right|} \sum_{n \in \{n\}_a^{(t)}} \mathrm{sim}(n^*, n) \tag{6.2}$$

The similarity function $\mathrm{sim}(\cdot, \cdot)$ measures how similar two nouns are and is defined as the exponentiated negative distance in semantic space which assigns differential weights to exemplars based on their relative distances to the query (higher similarities for more proximal exemplars):

$$\mathrm{sim}(n^*, n) = \exp\left(-\frac{d(n^*, n)^2}{h}\right) \tag{6.3}$$

$d(\cdot, \cdot)$ measures the Euclidean distance between nouns and h is a kernel parameter that we learn from data. The choice of the exemplar model and the similarity formulation is grounded in work on Generalized Context Model (Nosofsky 1986), which has recently been shown to predict the historical extension of Chinese numeral classifiers (Habibi et al. 2020). Here we examine whether the same exemplar-based processes of chaining might explain historical adjective extension. This model is also equivalent to performing kernel density estimation in semantic space defined by the likelihood function, and thus we use a kernel parameter h in the similarity function and normalize the term by dividing the resulting sum by h.

3.2.2 Prototype model

Motivated by earlier psychological work on prototype theory (Rosch & Mervis 1975) and related recent work on few-shot learning (Snell et al. 2017), we consider an alternative view of chaining based on category prototypes. Each adjective a is represented by a prototype at time t that captures the "gist" of noun referents for that category. We operationalize the prototype as the expectation of all exemplars within a category:

$$\vec{\mathrm{p}}_a = \mathbb{E}\left[n \in \{n\}_a^{(t)}\right] = \frac{1}{\left|\{n\}_a^{(t)}\right|} \sum_{n \in \{n\}_a^{(t)}} n \tag{6.4}$$

The likelihood function postulates a chaining mechanism that links a query noun to the adjective that has the closest prototype in semantic space:

$$p(n^*|a)^{(t)} \propto \mathrm{sim}(n^*, \vec{\mathrm{p}}_a) = \exp\left(-\frac{d(n^*, \vec{\mathrm{p}}_a)^2}{h'}\right) \tag{6.5}$$

Similar to the exemplar model, we use a kernel parameter h' that controls how quickly similarity scales with respect to the semantic distance between the query noun and the prototype. This model can behave differently from the exemplar model of chaining: even if a query noun is closer to the prototype of one adjective over an alternative adjective, a small set of exemplars closest to that noun can pull the query item to the alternative category (see Habibi et al. 2020 for a simulation that compares the properties of the exemplar and prototype models of chaining).

We also consider a variant of the prototype model in which the prototype representation for each adjective category remains static over time. That is,

$$\vec{\mathrm{p}}_a = \vec{\mathrm{p}}_a^{(t_0)}$$

for all $t > t_0$ where t_0 is the initial time of investigation. We refer to this variant as the progenitor model.

3.2.3 *k*-nearest neighbours model

In addition to the exemplar and prototype models, we consider a family of models based on k-nearest neighbours (k-NN). In a Bayesian framework, the k-NN likelihood of n^* pairing up with adjective a is proportional to whether its k closest neighbours $n_1, \ldots, n_k$ previously paired up with a, and inversely proportional to the size of category a:

$$p(n^*|a)^{(t)} \propto \frac{1}{\left|\{n^*\}_a^{(t)}\right|} \sum_{j=1}^{k} I(n_j \in \{n\}_a^{(t)}) \tag{6.6}$$

Here the sum is over the k nouns closest to n^* in semantic space. When this likelihood is combined with the prior, the k-NN posterior probability amounts to n^*'s k closest neighbours voting for each of the adjectives that they previously paired up with.

This formulation of k-NN can be viewed as a "hard version" of the exemplar model where k is a discrete analog of the kernel parameter h. We report $k = 1$ and $k = 10$ in our experiments.

3.3 Prior distribution

We formulate a type-based prior $p\left(a\right)^{(t)}$ which specifies how likely adjective a is to be paired with any noun based on the set size of its noun referents at time t. This prior formulation predicts that a's probability of appearing in a novel adjective-noun pairing is directly proportional to the number of unique nouns it has previously paired up with:

$$p(a)^{(t)} = \frac{\left|\{n\}_a^{(t)}\right|}{\sum_{a' \in \mathcal{A}} \left|\{n\}_{a'}^{(t)}\right|} \tag{6.7}$$

The rationale behind this choice of prior is as follows: if semantic chaining underlies the emergence of novel adjective-noun pairs, then adjectives that have paired with more nouns would have a higher a priori probability of attracting a query noun n^* via linking it to semantically similar nouns which are more likely to have previously co-occurred with a (Luo & Xu 2018). This rich-get-richer process is also supported by work on how semantic networks grow through preferential attachment (Steyvers & Tenenbaum 2005).

This category-size-based prior serves as our baseline model when making adjective predictions for n^* at time $t + \Delta$, where $p\left(a|n^*\right)^{(t+\Delta)} = p\left(a\right)^{(t)}$. We focus on the type-based representation as opposed to token frequencies because work from Habibi et al. (2020) has shown that a type-based prior worked better than a token-based prior in predicting the extension of grammatical categories.

3.4 Semantic space

To construct a semantic space for the nouns, we use word embeddings, particularly Word2Vec, commonly used for distributed semantic representation in natural language processing (Mikolov et al. 2013). We choose this construction of semantic space partly because it has been demonstrated to be effective in predicting grammatical category extension (Habibi et al. 2020). However, adjective usage is likely to entail a semantic representation richer than purely linguistic information, and future work should explore alternative methods for constructing semantic space such as those based on perceptual features and lexical taxonomic structures.

Since the word co-occurrence distributions are constantly changing over time, our semantic representations (of nouns) also need to be updated accordingly. For this reason, we use diachronic (or historical) Word2Vec embeddings (Hamilton et al. 2016) where at each time t, the embedding for a noun is based on its co-occurrence profiles at time t, relatively independent to future co-occurrences. In this respect, the predictions made by our models are in some sense "zero-shot", or deprived of semantic information into the future.

4 Data

We extracted a large database of historical adjective-noun pairings over the past 150 years (1850–2000). We collected these data from the Google Books corpus (Michel et al. 2011) which contains sentence fragments from historical books over the past five centuries. Within Google Books, the English All (EngAll) corpus accounts for 8.5×10^{11} tokens and roughly 4% of all books ever published. The diversity and size of the EngAll corpus should reflect how the English language has been used over the past centuries, which makes our adjective-noun co-occurrence dataset suitable for evaluating hypotheses about chaining.

We collected adjective-noun co-occurrence counts from the EngAll corpus. First, we extracted all bigrams from the EngAll corpus in which the first token is an adjective and the second is a noun (by part-of-speech tags specified in the data) along with the corresponding timestamp. Since the corpus is likely to contain noise, we standardized the set of nouns and adjectives by only considering those present in WordNet (Miller 1995), which yields approximately 67k nouns and 14k adjectives.

We collapsed raw co-occurrence counts into decadal bins by choosing $\Delta = 10$ years. This yielded our adjective-noun pairings dataset which consists of entries of the form (a, n, count, t). In each decade t, we used a Word2Vec language model pre-trained on historical text (i.e., digitized books from Google Books) for the semantic representation. For our analyses, we worked with a subset of the collected data (discussed in the next section), due to both considerations of sampling diversity and computational feasibility. To construct semantic representations across decades, we used diachronic Word2Vec embeddings which were trained using the EngAll corpus. (Hamilton et al. 2016) also chose to construct diachronic Word2Vec embeddings decade-by-decade for similar reasons.

We now describe three adjective sets $\mathcal{A}$. The purpose of evaluating our models on three different adjective sets is to obtain representative samples of the adjectives, and to ensure our hypotheses are robust to the choice of adjectives.

(1) *Frequent adjectives.* We use multiple ways to construct $\mathcal{A}$ such that it covers a broad scope and show our results are reproducible and agnostic to choice of adjectives. To construct a set of 200 adjectives that cover a broad range of descriptions, we first collected word vectors of all adjectives in the Google Books corpus using a pre-trained Word2Vec model. Next, we clustered the adjectives into 20 clusters and picked 10 adjectives from each to construct our set $\mathcal{A}$ of 200 adjectives. We applied this clustering procedure to obtain a feasibly large yet diverse set of adjectives for the analyses, and we used the k-means algorithm for clustering. Adjectives were sampled from each cluster based on their usage frequencies, and only considered against other adjectives within the same cluster during sampling. We refer to this set as Frq-200, with examples shown in Table 6.1.

(2) *Random adjectives.* To ensure that the sampling scheme for choosing $\mathcal{A}$ is not biased towards token frequencies, we also constructed another set of 200 adjectives by repeating the clustering step described above, but we replaced frequency sampling with uniform sampling. We refer to this dataset as Rand-200. As Table 6.1 shows, adjectives drawn from the same cluster are semantically similar between Frq-200 and Rand-200, but less common in the latter set.

(3) *Synaesthetic adjectives.* We also consider the third set of *synaesthetic adjectives* (Syn-65) defined by Williams (1976), as a more focused domain that is known to undergo semantic change. This set includes 65 adjectives that exhibit regular semantic shift historically. We will refer to this set as Syn-65.[3]

Data and code from our analyses are available at https://git.io/JqeyK.

5 Results

We present results in two steps. First, we examine the set of chaining algorithms described on novel adjective-noun pairings that appeared during 1850–2000, and we evaluate whether the exemplar model would better predict these data than the alternative models. Second, we perform a more focused analysis to examine whether the chaining algorithms predict extensional patterns in adjectives that show most and least semantic change over the past 150 years.

[3] There are in fact 64 unique adjectives in this set and WordNet captures 61 of these adjectives. See Williams (1976) for a comprehensive list.

Table 6.1: A comparison of some adjectives in Frq-200 and Rand-200 grouped according to the cluster they were drawn from. Notice that the clusters (per column) align semantically, however the adjectives in Frq-200 are more frequently represented in the English lexicon than those in Rand-200.

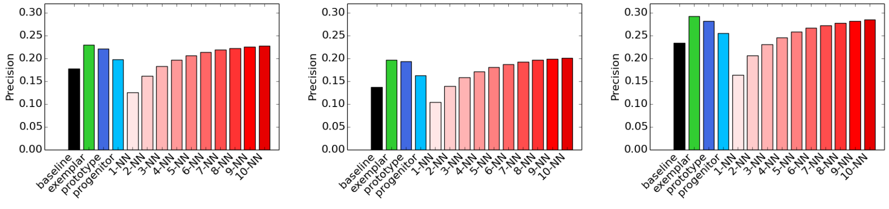
Frq-200	Rand-200	Frq-200	Rand-200
Asian	*Hungarian*	*polite*	*chatty*
Christian	*Thai*	*intelligent*	*unorthodox*
American	*Cornish*	*passionate*	*amiable*
European	*Catalan*	*energetic*	*communicative*

5.1 Evaluation of chaining algorithms

We evaluated the set of chaining algorithms on their ability to predict which adjectives $a \in \mathcal{A}$ would pair up with a given noun n^* in decade $t + \Delta$ given information about n^* up to and including decade $t > t_0$, where t_0 is the base decade. This information includes co-occurrences between all nouns n and adjectives $a \in \mathcal{A}$ at or before decade t, as well as time-dependent word embeddings at each decade taken from Hamilton et al. (2016). We chose t_0 as the 1840s and built a base lexicon from adjective-noun co-occurrences between t_0 and the 2000s. The 1860s was the first decade for which we report model prediction, and we used the 1850s as our "training decade" to estimate the kernel parameters for the exemplar and prototype models.

We define pairings (a, n^*) to be novel in decade $t + \Delta$ if and only if (i) a co-occurred with n^* in decade $t + \Delta$ beyond a certain threshold (which we set to 2), and (ii) a never appeared with n^* beyond that threshold in any decade $t' < t$. Using these criteria allowed us to eliminate noise from co-occurrence statistics. Given a noun n^*, each model's output was a categorical distribution $p\left(a|n^*\right)^{(t+\Delta)}$ over all adjectives $a \in \mathcal{A}$. The model was then scored on its precision accuracy on the set of adjectives that first co-occurred with n^* in decade $t + \Delta$. That is, if n^* co-occurred with m new adjectives in $\mathcal{A}$ in decade $t + \Delta$, we took the top m adjectives with the highest posterior probabilities that had not previously co-occurred with n^* as the set of retrieved positives. This evaluation metric calculates the percentage of correct predictions a model makes, and it is identical to the metric used in previous work for the prediction of historical extensions of classifier categories (Habibi et al. 2020). We report the total precision for all models and use this metric as an objective function to learn the kernel parameters from the initial training decade. We consider two types of predictive tasks when making predictions for noun n^* in decade t: taking as ground truth adjectives that co-occur with n^* (1) specifically in the immediate future decade $t + \Delta$, and (2) all future decades $t' > t$ up to the terminal decade 1990s.

We summarize results from our experiments for the three differently sampled adjective sets $\mathcal{A}$. As Figure 6.4 shows, the exemplar model has the highest predictive performance, followed closely by the 10-NN and prototype models. The exemplar, prototype, and 10-NN models perform substantially better than the baseline. These results provide evidence that chaining may rely on mechanisms sensitive to semantic neighbourhood density, best captured by the exemplar model. We also observed that the 10-NN model did not perform better than the exemplar model as the kernel parameter is a continuous analog of k and is optimized for precision, but increasing k in k-NN beyond 10 did help to improve model prediction suggesting that local neighbourhood density matters in predicting adjective extension. The progenitor model, a variant of the prototype model with static prototypes determined in decade t_0, is considerably worse than the prototype model with a moving prototype. This relationship between the prototype and progenitor models that we observe indicates that if the prototype model is the closest underpinning of adjective extension, then $\{n\}_a^{(t)}$ largely influences which nouns adjective a will extend to and that each adjective category "center" updates once novel adjective-noun pairings are formed. We also observed that the baseline or prior model performed worse than the exemplar and prototype models, suggesting that semantic relations matter in adjective-noun pairing, above and beyond the size-based adjective priors.



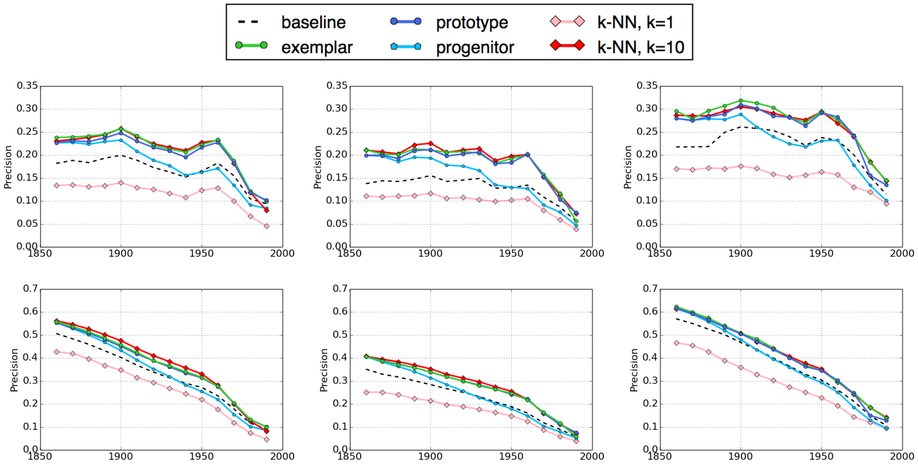
(a) Frq-200

(b) Rand-200

(c) Syn-65

Figure 6.4: Aggregate precision accuracy for all models (including k-NN from $k = 1$ to $k = 10$) across all time periods on each of our three adjectives sets.

Further results with year-over-year accuracy breakdowns are shown in Figure 6.5. The predictive accuracy falls in later decades since there are fewer novel adjective-noun pairings to predict. Our results hold generally across the three adjective sets, and they suggest that semantic neighbourhood density is an important factor contributing towards adjective extension as the exemplar and 10-NN



(a) Frq-200

(b) Rand-200

(c) Syn-65

Figure 6.5: Model predictive accuracy on the Frq-200, Rand-200, and Syn-65 adjective sets. *Top row*: Predictive accuracy when only novel adjective-noun pairs in the following decade are considered. *Bottom row*: Predictive accuracy when all future adjective extensions are considered.

models achieve overall better predictive accuracy over the other alternative models.

Table 6.2 provides some examples of model prediction and highlights the limitations of the approach. It is worth noting that while the exemplar model performed well in comparison to the other models, all models failed to predict parts of the empirical data. This issue might be partly due to the fact that our semantic representation of nouns is inadequate to capture the kinds of rich knowledge that determines adjective modification of nouns, and partly due to the historical events that add randomness to the process, e.g., how alcohol prohibition in the 1920s made *illegal* an appropriate adjective modifier for *alcohol*, and how *American* and *Vietnam* became associated in context presumably due to the Vietnam War around the 1960s.

5.2 Chaining in semantically changing and stable adjectives

We next examine the extent to which the chaining algorithms predict extensional patterns in both semantically changing and stable adjectives in history. Because the chaining view presumes meaning change to take incremental (as opposed to abrupt) steps, it is plausible that it is less effective in predicting adjective extension in those adjectives that show substantial change in meaning over time.

Table 6.2: Examples of model prediction on the Frq-200 adjective set. Adjectives with an asterisk (*) indicate true positives retrieved by models. We present predictions for nouns _cigarette_, _alcohol_, and _Vietnam_ as the adjectives they first pair with in the 1880s, 1920s, and 1960s respectively reflect sentiment (e.g., _social cigarette_) or historic events (e.g., _illegal alcohol_ due to prohibition, _American Vietnam_ due to the Vietnam war).

noun & decade	_cigarette_, 1880s
new adjectives	_better_, _modern_, _several_, _excessive_, _American_, _social_
baseline prediction	_original_, _particular_, _English_, _natural_, _perfect_, _modern_* (1/6)
exemplar prediction	_black_, _red_, _English_, _poor_, _original_, _particular_ (0/6)
prototype prediction	_red_, _black_, _dry_, _warm_, _cold_, _English_ (0/6)
10-NN prediction	_original_, _warm_, _particular_, _red_, _English_, _dry_ (0/6)
noun & decade	_alcohol_, 1920s
new adjectives	_female_, _analogous_, _red_, _bitter_, _marked_, _illegal_
baseline prediction	_perfect_, _extraordinary_, _moral_, _physical_, _western_, _christian_ (0/6)
exemplar prediction	_red_*, _moral_, _artificial_, _dense_, _perfect_, _marked_* (2/6)
prototype prediction	_artificial_, _perfect_, _marked_*, _red_*, _physical_, _moral_ (2/6)
10-NN prediction	_red_*, _moral_, _dense_, _perfect_, _analogous_*, _artificial_ (2/6)
noun & decade	_Vietnam_, 1960s
new adjectives	_western_, _tropical_, _eastern_, _colonial_, _particular_, _more_, _top_, _poor_, _American_
baseline prediction	_same_, _more_*, _great_, _particular_*, _American_*, _different_, _natural_, _human_, _English_ (3/9)
exemplar prediction	_western_*, _eastern_*, _more_*, _particular_*, _great_, _colonial_*, _inner_, _same_, _poor_* (6/9)
prototype prediction	_great_, _same_, _western_*, _more_*, _American_*, _eastern_*, _particular_*, _European_, _French_ (5/9)
10-NN prediction	_western_*, _eastern_*, _more_*, _tropical_*, _colonial_*, _great_, _better_, _inner_, _particular_* (6/9)

However, if chaining reflects a generic mechanism of meaning change, we should expect the models described to predict both semantically changing and stable adjectives.

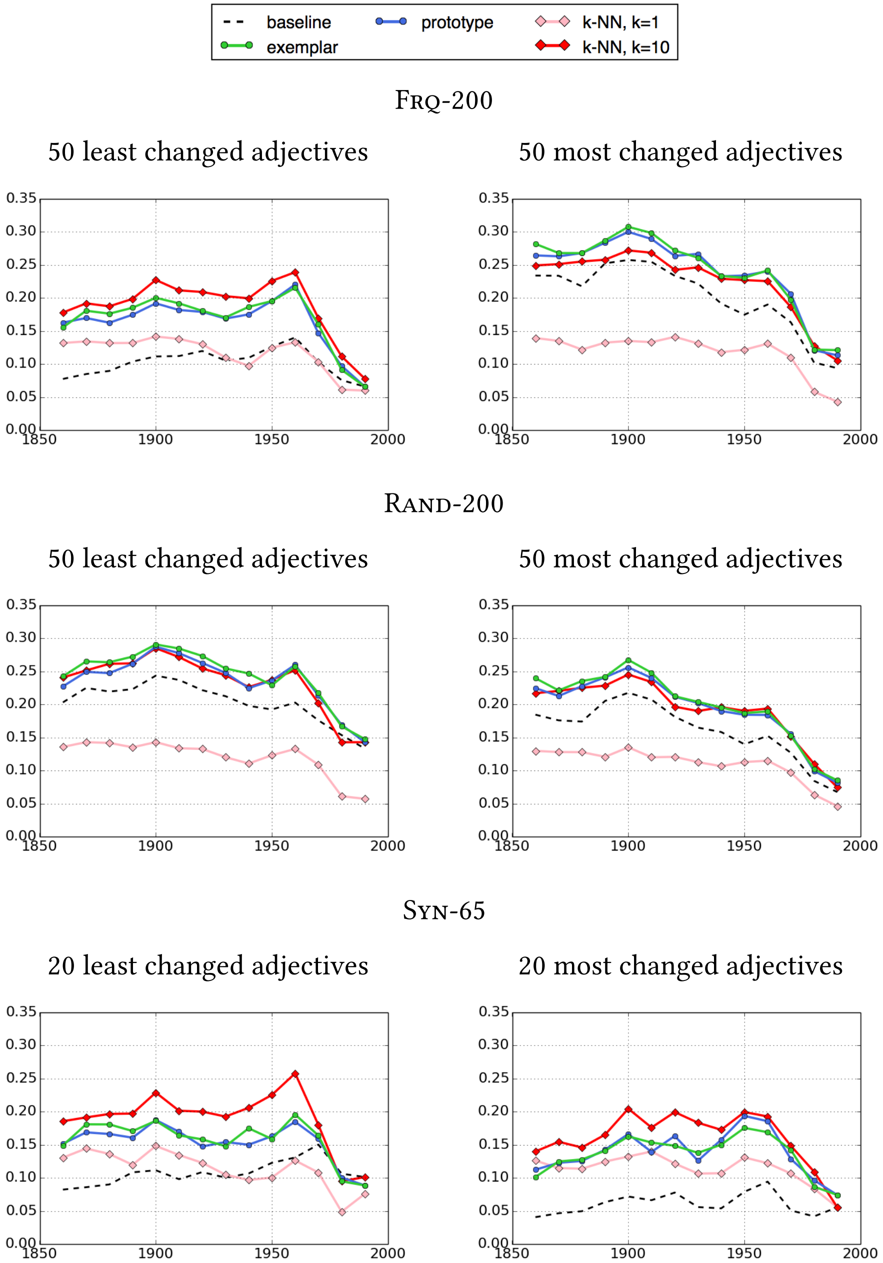
To investigate this issue, we performed a group comparison where we split each adjective set into two subsets: a semantically changing group that showed the highest degrees of semantic change, and a semantically stable group that showed the least degrees of semantic change. We defined the degree of semantic change of an adjective based on its semantic neighbourhood profiles during the flanking decades: 1850s and 1990s. We followed the same procedure as Xu et al. (2015), where we calculated the degree of overlap in 100 semantic neighbours in adjectives (using diachronic word embeddings from Hamilton et al. 2016) between the flanking decades and took the inverse of that quantity as degree of change: a fully stable adjective would have 100% overlap in its neighbourhood, whereas a highly changing adjective would have low % overlap in its neighbourhood. We then applied the same models of chaining to these two subgroups in each of the three adjective sets.[4]

We analyzed the 50 most and least changing adjectives from the Frq-200 and Rand-200 sets, and only 20 most and least changing adjectives from the Syn-65 set because it contained 61 adjectives in total. The results appear in Figure 6.6. We observed that the proposed algorithms of chaining, particularly the exemplar, prototype, and 10-NN models, perform substantially better than the frequency baseline. This observation holds for both the semantically changing and stable adjective subgroups, suggesting that chaining mechanisms apply equally to these adjective sets. In both the Frq-200 and Rand-200 sets, the exemplar model consistently outperforms the alternative models in predictive accuracy over time, yet its performance is not the strongest in the Syn-65 (though this particular set has the smallest subset size of 20 adjectives). These results suggest that chaining is a generic mechanism in historical adjective extension.

6 Discussion

Our findings support the overall hypothesis that semantic neighbourhood density influences how novel adjective-noun pairings emerge, although the distinction between the exemplar model and the alternative models is small for drawing strong conclusions from this initial investigation. Nevertheless, all the models we

[4] For the prototype model, we only present results based on the (moving) prototype model because it was shown to be a superior model than the progenitor model that assumes the prototype to be time-invariant, both in Section 5.1 and Habibi et al. (2020).



FRQ-200

50 least changed adjectives

50 most changed adjectives

RAND-200

50 least changed adjectives

50 most changed adjectives

SYN-65

20 least changed adjectives

20 most changed adjectives

Figure 6.6: Model predictive accuracy on the most semantically changing and stable adjectives from FRQ-200, RAND-200, and SYN-65 adjective sets.

examined perform considerably better than the baseline model. Our work mirrors existing studies on chaining in the extension of grammatical categories (Lakoff 1987, Bybee et al. 1994, Habibi et al. 2020), and we discuss its limitations and implications toward a general theory of word meaning extension.

6.1 Limitations

Our formulation of chaining depends on semantic similarity. One drawback of this assumption is that although chaining mechanisms may retrieve nouns that are similar to a query noun, there is no independent mechanism of checking whether the adjective-noun pairing is plausible. That is, our implementation of chaining does not explicitly "perform a check" as to whether a predicted adjective-noun pairing is sensible. As adjectives accumulate novel senses, the set of possible nouns they can pair with will also vary due to external factors orthogonal to the internal mechanism of chaining. Here we acknowledge this limitation and consider it an important future direction to explore the interaction of internal and external factors that co-shape word meaning extension and semantic change.

Throughout our analyses we have assumed that distributed semantic representations, or word embeddings, are sufficient to capture the meaning of nouns. In particular, we used Word2Vec to capture distributional meaning of words from linguistic context, but other variants of semantic representation are available and should be considered in future explorations. Importantly, perceptual (e.g., visual) features might be especially relevant for constructing the meaning of concrete nouns, and our current construction of the semantic space might not capture these features. There exists computational work that explores adjective meaning using a combination of visual and linguistic information. For instance, Lazaridou et al. (2015) applied cross-modal mappings between visual and linguistic representations to assign adjective labels to visual inputs, and Nagarajan & Grauman (2018) followed up by learning a linear mapping that predicts adjective descriptors based on visual input. However, one limiting factor of these cross-modal approaches is that they may not be relevant to predicting adjective pairings with abstract nouns where perceptual grounding is more difficult to establish. In these cases, both socio-cultural factors and cognitive devices such as metaphor may be relevant in predicting adjective extension, above and beyond the semantic representation and the simple chaining mechanisms that we have considered.

Our analyses have relied on written text (i.e., books) which might not be fully representative of natural language use that also involves colloquialism and conversations (represented more accurately in spoken text corpora). The interpretations we drew from our analyses are thus restricted to formal forms of language, although they are also useful reflections of conventional language use. Earlier work by Williams (1976) on synaesthetic adjectives has also used dictionaries as a source of investigation, and a potential research direction is to examine the properties of adjective meaning extension or change in both written and spoken text. Written language is likely to be a delayed reflection of spoken language, and as such we might expect changes in word meaning and usage in spoken language to precede those in written text. Colloquialism may also add nuances beyond this difference, whereby language use is notably more casual and flexible (partly due to the socio-cultural knowledge involved), e.g., emergent adjective usages in slang might be harder to predict in comparison to the case of formal written text.

6.2 Relations of chaining and semantic change

The proposal of semantic chaining as initially described by Lakoff (1987) has focused on the formation of complex linguistic categories, particularly grammatical classes such as classifiers and prepositions. Although Lakoff did not discuss extensively the relations of chaining and historical semantic change, the anecdotal cases that he described have assumed a connection between the chaining mechanism and the process of polysemy, or word sense extension. For instance, in both of his accounts on the extension of classifier systems and spatial prepositions, he described how polysemous extensions – e.g., how Dyirbal classifiers group ideas related to women, fire, and dangerous things (Lakoff 1987), and how English *over* expresses a broad range of spatial configurations and metaphorical senses (Lakoff 1987, Brugman 1988) – might depend on image schematic transformations that are reflected through a process of chaining where one referent or sense links to another in complex chain-like structures. Recent computational work has extended these ideas in a formal setting and found that models of chaining – similar to those described in this chapter – can explain historical word sense extension in the English lexicon (Ramiro et al. 2018), although such models are far from perfect.

A caveat in both that study and Lakoff's (1987) work is the under-specification of the diverse knowledge involved in word meaning extension and semantic change, which is clearly beyond the embedding-based semantic representation presented here. In this respect, whether or how the theory of chaining can explain the diverse range of semantic change in adjectives and other word classes remains an important open question.

6.3 Toward a general theory of word meaning extension

The lexicon is an innovative product of the mind, and here we have focused on examining one critical form of lexical innovation that involves word meaning extension. A general account of word meaning extension in natural language ought to explain how it functions at different temporal scales not restricted to a historical setting.

There are at least three levels at which word meaning extension can occur, summarized in decreasing temporal scales: (1) across languages, the relics of word meaning extension are reflected in the colexification and polysemy structures that are likely a result of language evolution through tens of thousands of years (François 2008, Youn et al. 2016), e.g., how a single word form like *fire* can denote the senses of 'physical fire', 'flame', and 'anger'; (2) within a language, word meaning extension can occur in language change during hundreds of years (Sweetser 1991, Traugott & Dasher 2001), e.g., how words like *mouse* originally referred to 'a type of rodent' later extended to express 'a computer device'; (3) in child development typically within the first 2–3 years of life, children extend word meaning toward novel objects for which they lack the proper words in the form of overextension (Vygotsky 1962, Clark 1978, Rescorla 1980), e.g., how children use *ball* to refer to 'a balloon'. Characterizing the common mechanisms and knowledge underlying these phenomena will shed light on word meaning extension as a general strategy for making innovative use of a finite lexicon.

Recent studies have made initial progress toward this direction. For instance, Ferreira Pinto Jr. & Xu (2019) developed a multi-modal semantic framework based on the exemplar model of chaining and showed that it predicts children's overextension behavior in a variety of studies from the psychological and linguistic literature. Xu et al. (2020) showed that the frequency variation in cross-linguistic colexification, i.e., why certain senses are more commonly grouped (e.g., 'fire'–'flame') under a single word form than others (e.g., 'fire'–'anger') can be explained by a principle of cognitive economy, whereby senses that are frequently colexified across languages tend to be easily associable – an argument that is consistent with the chaining account presented here. However, there is a critical lack of demonstrating how the approaches and principles identified in language development and cross-linguistic settings can also explain historical semantic change attested in the world's languages.

We believe that a general formal account of word meaning extension will involve three key ingredients related to the chaining processes discussed in this chapter: (1) algorithmic formulations such as the exemplar model described that capture the mechanisms of semantic chaining; (2) rich knowledge structures that support these mechanisms toward a diverse range of extensional strategies such as metonymy and metaphor; (3) external socio-cultural influences or events that provide the driving force for word meaning extension.

7 Conclusion

We have presented a computational approach to explore regularities in the historical composition of adjectives and nouns through probabilistic models of chaining. Our approach provides clues to the generative mechanisms that give rise to novel adjective usages over time, and we hope it will stimulate future work on the semantic representation and the interaction of cognitive and socio-cultural underpinnings of word meaning extension and semantic change.

Acknowledgements

We thank Barend Beekhuizen, Lars Borin, Sammy Floyd, Adam Jatowt, Charles Kemp, and Nina Tahmasebi for helpful discussion and constructive comments. We also thank Amir Ahmad Habibi for sharing data and code. KG is supported by a Bell Graduate Scholarship and a Vector Scholarship in Artificial Intelligence. YX is funded through an NSERC Discovery Grant RGPIN-2018-05872 and a SSHRC Insight Grant #435190272.

Abbreviations

NN	nearest-neighbour

Appendix A Adjective sets

Here we present all adjectives used in our analysis, namely from the Frq-200, Rand-200, and Syn-65 adjective sets. Adjectives with an asterisk (*) are included in at least two of the three adjective sets. The first table gives the adjectives that constitute Syn-65, and we note two important details about this set. First, the set of synaesthetic adjectives proposed by Williams (1976) actually contains 64 unique adjectives as *light* is repeated. Second, the Google Books corpus ties all tokens to words in WordNet, and since *acrid*, *aspre*, and *tart* (all synaesthetic adjectives) are not WordNet adjectives, we could not reliably measure their uses through time. For this reason, we excluded these from Syn-65 and have 61 adjectives in total, listed in Table 6.3.

Table 6.3: List of 61 adjectives in Syn-65. Adjectives with an asterisk (*) appear in at least one of Frq-200 and Rand-200 as well.

Syn-65						
acute	*cloying*	*dulcet**	*grave*	*light*	*quiet*	*sour*
austere	*coarse*	*dull*	*hard*	*little*	*rough*	*strident*
big	*cold**	*eager*	*harsh*	*loud*	*shallow*	*sweet*
*bitter**	*cool*	*empty*	*heavy*	*low*	*sharp*	*thick*
bland	*crisp*	*even*	*high*	*mellow*	*shrill*	*thin*
bright	*dark*	*faint*	*hollow*	*mild*	*small*	*vivid*
*brilliant**	*deep*	*fat*	*hot*	*piquant*	*smart*	*warm**
brisk	*dim*	*flat*	*keen*	*poignant*	*smooth*	
clear	*dry**	*full*	*level*	*pungent*	*soft*	

Next, we present the Frq-200 and Rand-200 adjective sets and the clusters we used for the analysis, listed in Table 6.4. Since these two sets draw adjectives from identical clusters, we present the two adjective sets so we can easily compare adjectives drawn from same cluster between the two sets.

Table 6.4: Lists of adjectives and clusters in Frq-200 and Rand-200.

cluster 1 of 20		cluster 2 of 20		cluster 3 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
casual	*amiable*	*bare*	*contorted*	*sufficient*	*alterable*
eccentric	*chatty*	*curly*	*dainty*	*analogous*	*contemporaneous*
energetic	*communicative*	*eyed*	*furrowed*	*equal*	*reconcilable*
entertaining	*fiery*	*female*	*hale*	*calculable*	*chargeable*
enthusiastic	*fluent*	*feminine*	*horny**	*receivable*	*distributive*
forgiving	*guileless*	*horny**	*limber*	*derived*	*accessary*
glib	*lovable*	*male*	*sage*	*binding*	*lineal*
intelligent	*loyal*	*naked*	*skeletal**	*indirect*	*allotted*
passionate	*patriotic*	*pale*	*smoky*	*undivided*	*noncommercial*
polite	*unorthodox*	*skeletal**	*swaggering*	*eligible*	*classifiable*
cluster 4 of 20		cluster 5 of 20		cluster 6 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
*cold**	*chilly*	*algebraic*	*binary*	*blind*	*intact*
dense	*cold**	*conventional*	*biotic*	*impossible*	*irretrievable*
*dry**	*drizzling*	*discrete*	*crystalline*	*incomplete*	*malfunctioning*
eastern	*encroaching*	*electrical*	*fusible*	*isolated*	*obscure*
hardy	*fertile*	*microscopic*	*geometric*	*pregnant*	*overlooked*
northern	*funicular*	*multicellular*	*interfacial*	*scarce*	*powerless*
south	*homeward*	*predictive*	*modular*	*silent*	*unmarked*
tropical	*littoral*	*rotational*	*perceptual*	*submerged*	*unstable*
*warm**	*unincorporated*	*thermal*	*refrigerant*	*unknown*	*unstudied*
western	*watery*	*volcanic*	*stratified*	*unrelated*	*valueless*

cluster 7 of 20		cluster 8 of 20		cluster 9 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
appropriate	*complex*	*alien*	*antipodal*	*everyday*	*approaching*
balanced	*delighted*	*colonial*	*congruous*	*firm*	*descending*
basic	*foolproof*	*divine*	*dynastic*	*more*	*fiddling*
better	*grateful*	*heavenly*	*hierarchical*	*original*	*former*
different	*intensive*	*human*	*invariable*	*particular*	*intensifying*
natural	*knowledgeable*	*inner*	*overt*	*physical*	*probable*
positive	*livable*	*medieval*	*paschal*	*preliminary*	*rental*
solid	*realistic*	*modern*	*protestant*	*same*	*reverse*
superior	*structured*	*moral*	*recessive*	*several*	*sliding*
sure	*varied*	*philosophical*	*sacred*	*top*	*thirteenth*

cluster 10 of 20		cluster 11 of 20		cluster 12 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
allergic	*carcinogenic*	*black*	*ceramic*	*bent*	*hysterical*
antibiotic	*coagulate*	*circular*	*cyclopean*	*bourgeois*	*inattentive*
artificial	*colorless*	*concave*	*fireproof*	*corrupt*	*irreligious*
dietary	*milky*	*crimson*	*legible*	*disreputable*	*lunatic*
fibrous	*nonfat*	*distinctive*	*rectilinear*	*domineering*	*opportunist*
liquid	*pulpy*	*fluorescent*	*sleek*	*evil*	*parochial*
mucous	*scented*	*incised*	*tucked*	*fascist*	*possessive*
powdery	*spongy*	*red*	*umber*	*jugular*	*resentful*
raw	*steamed*	*tubular*	*unglazed*	*pious*	*uncongenial*
synthetic	*vanilla*	*white*	*Venetian*	*warlike*	*unengaged*

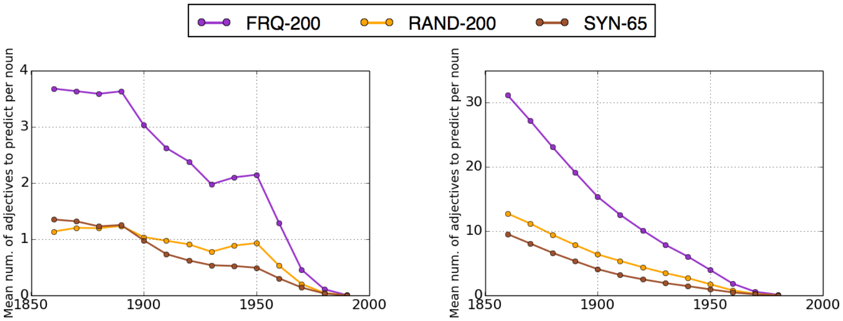
cluster 13 of 20		cluster 14 of 20		cluster 15 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
*bitter**	*brokenhearted*	*affected*	*bottomed*	*abusive*	*appalling*
debilitating	*confused*	*buried*	*credited*	*deplorable*	*bias*
emotional	*delirious*	*distributed*	*jammed*	*exaggerated*	*capricious*
hopeless	*disturbed*	*given*	*owned*	*excessive*	*exorbitant*
*odd**	*odd**	*left*	*rose*	*illegal*	*hostile*
poor	*patchy*	*marked*	*scattered**	*simplistic*	*imprecise*
troubled	*regretful*	*modified*	*settled*	*undue*	*inelegant*
*unhappy**	*thirsty*	*scattered**	*shattered*	*unintentional*	*innocuous*
weird	*unhappy**	*used*	*surrounded*	*unproductive*	*unbalanced*
worst	*untidy*	*worn*	*sworn*	*wrong*	*unsound*

cluster 16 of 20		cluster 17 of 20		cluster 18 of 20	
Frq-200	Rand-200	Frq-200	Rand-200	Frq-200	Rand-200
adrenal	*cesarean*	*American*	*Arabian*	*brilliant**	*adored*
alveolar	*endoscopic*	*Asian*	*Catalan*	*conspicuous*	*commanding*
bivariate	*hemorrhagic*	*Christian*	*Chinese*	*ecstatic*	*fantastic*
cardiovascular	*hyoid*	*Dutch**	*Cornish*	*extraordinary*	*favorite*
clinical	*intervertebral*	*English*	*Dutch**	*fitting*	*gallant*
diagnostic	*lobular*	*European*	*Haitian*	*great*	*halcyon*
neural	*monovalent*	*French*	*Hungarian*	*incomparable*	*loved*
peritoneal	*normotensive*	*Roman*	*Kurdish*	*perfect*	*superb*
spinal	*valved*	*Serbian*	*Taiwanese*	*singular*	*tragic*
ulcerative	*vesicular*	*Spanish*	*Thai*	*startling*	*undefeated*

cluster 19 of 20		cluster 20 of 20	
FRQ-200	RAND-200	FRQ-200	RAND-200
budgetary	*agrarian*	*aesthetic*	*clarion*
civil	*catechetical*	*artistic*	*contemporary*
criminal	*clandestine*	*classical*	*darkling*
marital	*constitutional*	*clever*	*dulcet**
mental	*curricular*	*colloquial*	*earthy*
national	*hourly*	*dreamy*	*falsetto*
nuclear	*intramural*	*hilarious*	*longhand*
parental	*qualitative*	*intimate*	*ponderous*
regional	*recreational*	*narrative*	*soothing*
social	*sectional*	*rhetorical*	*wry*

Appendix B Temporal trends in model precision

As discussed in the main text, the model precision generally decreases across all models with time. As Figure 6.7 shows, the average number of nouns to predict in each decade decreases with time. This trend applies to both sets of true positives: only adjectives that first co-occur with a given noun n^* in decade $t + \Delta$, and also in any future decade. Consequently, the precision falls systematically in later decades because there are fewer novel pairings to predict in the data.



(a) In decade $t + \Delta$

(b) In any future decade $t' > t$

Figure 6.7: The average number of novel adjective-noun pairs remaining for each model to predict across all times and adjective sets. This value is computed across all nouns for which a predictive model makes adjective prediction.

Appendix C Semantically changing and stable adjectives

Table 6.5: Lists of most and least changed adjectives from the Frq-200, Rand-200, and Syn-65 sets, along with their top semantic neighbours during initial (1850s) and terminal (1990s) periods of investigation.

Frq-200	
Least Changed	Most Changed
1. *eccentric*	1. *classical*
1850s: *versatile*, *droll*, *impulsive*	1850s: *theological*, *modern*, *greek*
1990s: *incoherent*, *perverse*, *exquisite*	1990s: *greek*, *traditional*, *contemporary*
2. *casual*	2. *rhetorical*
1850s: *occasional*, *trivial*, *careless*	1850s: *idiomatic*, *didactic*, *fanciful*
1990s: *careless*, *informal*, *friendly*	1990s: *poetic*, *epistolary*, *grammatical*
3. *polite*	3. *colloquial*
1850s: *affable*, *hospitable*, *elegant*	1850s: *imaginative*, *analytic*, *bewitching*
1990s: *respectful*, *friendly*, *agreeable*	1990s: *epistolary*, *poetic*, *idiomatic*
4. *intelligent*	4. *narrative*
1850s: *honest*, *rational*, *inquisitive*	1850s: *detailed*, *circumstantial*, *brief*
1990s: *clever*, *energetic*, *minded*	1990s: *autobiographical*, *biblical*, *historical*
5. *enthusiastic*	5. *artistic*
1850s: *irrepressible*, *impulsive*, *passionate*	1850s: *scientific*, *architectural*, *literary*
1990s: *ardent*, *sincere*, *generous*	1990s: *intellectual*, *musical*, *poetic*
Rand-200	
Least Changed	Most Changed
1. *fluent*	1. *contemporary*
1850s: *versatile*, *idiomatic*, *sprightly*	1850s: *recorded*, *voluminous*, *anonymous*
1990s: *spoken*, *speaking*, *Arabic*	1990s: *literary*, *historical*, *classical*
2. *amiable*	2. *earthy*
1850s: *humane*, *affable*, *estimable*	1850s: *alkaline*, *gelatinous*, *nitrogenous*
1990s: *dignified*, *virtuous*, *pleasing*	1990s: *ceremonious*, *ravaging*, *disused*
3. *patriotic*	3. *soothing*
1850s: *loyal*, *disinterested*, *enlightened*	1850s: *melancholy*, *sweet*, *sympathetic*
1990s: *democratic*, *civic*, *loyal*	1990s: *calm*, *sweet*, *shrill*
4. *fiery*	4. *ponderous*
1850s: *fierce*, *resistless*, *malign*	1850s: *huge*, *cased*, *jingling*
1990s: *mutinous*, *treacherous*, *fierce*	1990s: *glistening*, *ethereal*, *noiseless*

5. *communicative* 1850s: *sociable, choleric, affable* 1990s: *symbolic, verbal, functional*	5. *clandestine* 1850s: *nefarious, illicit, adulterous* 1990s: *disfigured, patrician, sedate*

Syn-65

Least Changed	Most Changed
1. *bitter* 1850s: *astringent, sweet, poignant* 1990s: *sour, harsh, intense*	1. *shrill* 1850s: *blithe, deafening, inaudible* 1990s: *pitched, startled, muffled*
2. *bland* 1850s: *mild, unobtrusive, affable* 1990s: *unconverted, unadorned, affable*	2. *small* 1850s: *smaller, size, sized* 1990s: *sized, smaller, insignificant*
3. *coarse* 1850s: *dirty, threadbare, boned* 1990s: *thin, fine, stiff*	3. *mellow* 1850s: *lustrous, chilly, balmy* 1990s: *perfumed, fragrant, sportive*
4. *cold* 1850s: *clammy, wet, hot* 1990s: *warm, damp, windy*	4. *austere* 1850s: *unsocial, disdainful, rigid* 1990s: *matchless, apposite, erudite*
5. *cool* 1850s: *calm, chilly, warm* 1990s: *damp, hot, dry*	5. *pungent* 1850s: *juicy, ductile, astringent* 1990s: *mown, fresh, colorless*

References

Baroni, Marco & Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP 2010*. Cambridge.

Boleda, Gemma, Marco Baroni, Nghia Pham & Louise McNally. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS 2013*. Potsdam.

Brugman, Claudia Marlea. 1988. *The story of over: Polysemy, semantics, and the structure of the lexicon*. New York: Garland.

Bybee, Joan L. 2013. Usage-based theory and exemplar representations of constructions. In T. Hoffmann & G. Trousdale (eds.), *The Oxford handbook of construction grammar*. Oxford: Oxford University Press.

Bybee, Joan L., Revere Dale Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago: University of Chicago Press.

Clark, Eve V. 1978. Strategies for Communicating. *Child Development* 49. 953–959.

Ferreira Pinto Jr., Renato & Yang Xu. 2019. Children's overextension as communication by multimodal chaining. In *Proceedings of the forty-first Annual Meeting of the Cognitive Science Society*. Montréal.

François, Alexandre. 2008. Semantic maps and the typology of colexication: Intertwining polysemy networks across languages. In Martine Vanhove (ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations*, 163–215. Amsterdam: John Benjamins.

Grewal, Karan & Yang Xu. 2020. Chaining and historical adjective extension. In *Proceedings of the 42nd Annual meeting of the Cognitive Science Society*.

Habibi, Amir Ahmad, Charles Kemp & Yang Xu. 2020. Chaining and the growth of linguistic categories. *Cognition* 202. 104323.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Hilpert, Martin. 2007. Chained metonymies in lexicon and grammar. In Günter Radden, Klaus-Michael Köpcke, Thomas Berg & Peter Siemund (eds.), *Aspects of meaning construction*, 77–98. Amsterdam: John Benjamins.

Keuleers, Emmanuel. 2008. *Memory-based learning of inflectional morphology*. Universiteit Antwerpen. (Doctoral dissertation).

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lapata, Maria. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proceedings of NAACL 2001*, 63–70. ACL.

Lapata, Maria, Scott McDonald & Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Proceedings of EACL 1999*. Bergen: ACL.

Lazaridou, Angeliki, Georgiana Dinu, Adam Liska & Marco Baroni. 2015. From visual attributes to adjectives through decompositional distributional semantics. *Transactions of the ACL* 3. 183–196.

Luo, Yiwei & Yang Xu. 2018. Stability in the temporal dynamics of word meanings. In *Proceedings of the fortieth Annual Conference of the Cognitive Science Society*. Madison.

Malt, Barbara C., Steven A. Sloman, Silvia Gennari, Meiyi Shi & Yuan Wang. 1999. Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language* 40. 230–262.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331. 176–182.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger (eds.), *Advances in neural information processing systems*, 3111–3119. Red Hook, NY: Curran Associates, Inc. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11). 39–41.

Nagarajan, Tushar & Kristen Grauman. 2018. Attributes as operators: Factorizing unseen attribute-object compositions. In *Proceedings of the fifteenth European Conference on Computer Vision*. Munich.

Nosofsky, Robert M. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115. 39–57.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. *Typological Studies in Language* 45. 137–158.

Ramiro, Christian, Mahesh Srinivasan, Barbara C. Malt & Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences* 115. 2323–2328.

Ramsey, Rachel. 2017. *An exemplar-theoretic account of word senses*. Northumbria University. (Doctoral dissertation).

Rescorla, Leslie A. 1980. Overextension in Early Language Development. *Journal of Child Language* 7(2). 321–335.

Rosch, Eleanor & Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7. 573–605.

Schmidt, Lauren A., Charles Kemp & Joshua B. Tenenbaum. 2006. Nonsense and sensibility: Inferring unseen possibilities. In *Proceedings of the twenty-eighth Annual Conference of the Cognitive Science Society*. Vancouver.

Skousen, Royal. 1989. *Analogical modeling of language*. Dordrecht: Kluwer.

Sloman, Steven A., Barbara C. Malt & Arthur Fridman. 2001. Categorization versus similarity: The case of container names. In U. Hahn & M. Ramscar (eds.), *Similarity and categorization*, 73–86. New York: Oxford University Press.

Snell, Jake, Kevin Swersky & Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*. Long Beach.

Steyvers, Mark & Joshua B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science* 29. 41–78.

Sweetser, Eve. 1991. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge: Cambridge University Press.

Traugott, Elizabeth Closs & Richard B. Dasher. 2001. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Vecchi, Eva M., Marco Marelli, Roberto Zamparelli & Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science* 41(2). 102–136.

Vecchi, Eva M., Roberto Zamparelli & Marco Baroni. 2013. Studying the recursive behaviour of adjectival modification with compositional distributional semantics. In *Proceedings of EMNLP 2013*. Seattle: ACL.

Vygotsky, Lev Semenovich. 1962. *Language and thought*. Cambridge: MIT Press.

Williams, Joseph M. 1976. Synaesthetic Adjectives: A Possible Law of Semantic Change. *Language* 32. 461–78.

Xu, Kelvin, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Richard S. Zemel Ruslan Salakhutdinov & Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the thirty-second International Conference on Machine Learning*. Lille.

Xu, Yang, Khang Duong, Barbara C. Malt, Serena Jiang & Mahesh Srinivasan. 2020. Conceptual relations predict colexification across languages. *Cognition* 201. 104280.

Xu, Yang, Terry Regier & Barbara C. Malt. 2016. Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science* 40. 2081–2094.

Youn, Hyejin, Logan Sutton, Eric Smith, Cristopher Moore, Jon F. Wilkins, Ian Maddieson, William Croft & Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences* 113. 1766–1771.

Zanzotto, Fabio Massimo, Ioannis Korkontzelos, Francesca Fallucchi & Suresh Manandhar. 2010. Estimating linear models for compositional distributional semantics. In *Proceedings of COLING 2010*. Beijing: ACL.

Chapter 7

Cross-lingual laws of semantic change

Ana-Sabina Uban[a,b], Alina Maria Ciobanu[a] & Liviu P. Dinu[a]

[a]Human Language Technologies Research Center, University of Bucharest
[b]PRHLT Research Center, Universitat Politècnica de València

Semantic divergence in related languages is a key concern of historical linguistics. In order to complement existing research on semantic change, which in most previous approaches relies on the comparative analysis of diachronic corpora, we propose a novel approach for measuring semantic shifts synchronically. In this chapter, we investigate semantic change across languages by measuring the semantic distance of cognate sets using cross-lingual word embeddings. We define a measure of cognate divergence and show how it can be used as a measure of language semantic divergence. We propose an algorithm to detect and correct false friends as an application for natural language processing. We hypothesize that false friends fall on a spectrum, and define a corresponding notion of "falseness" and a methodology for its quantification. We evaluate the algorithm based on WordNet and on manually curated lists of true cognates and false friends, showing accuracy values exceeding 80%, and we show how choosing a falseness level as a threshold for detecting false friends can affect the performance of the detection algorithm. We further study the properties of the semantic divergence of cognates, and verify whether hypothesized laws of semantic change (namely the law of conformity and the law of innovation) hold in the multilingual setting, thus formulating the first laws of cross-lingual semantic change. We further study the mathematical relation between polysemy and frequency on the one hand, and falseness on the other hand, and identify polynomials that optimally model their relationship, leading to equations describing cross-lingual semantic change in relation to word properties.

1 Introduction

Semantic change – that is, change in the meaning of individual words (Campbell 1998) – is a continuous, inevitable process stemming from numerous reasons

Ana-Sabina Uban, Alina Maria Ciobanu & Liviu P. Dinu. 2021. Cross-lingual laws of semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 219–260. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040314

and influenced by various factors. Words are continuously changing, with new senses emerging all the time. Campbell (1998) presents no less than 11 types of semantic change that are generally classified in two broad categories: narrowing and widening. The author states that most linguists found structural and psychological factors to be the main cause of semantic change, but the evolution of technology and cultural and social changes are not to be omitted.

Intra-lingual semantic shift has been previously studied in computational linguistics, but monolingual studies can only provide a limited picture of the evolution of word meanings, which often develop in a multilingual setting, with new words entering the language through inheritance and borrowing. Measuring semantic divergence across languages can be useful in theoretical and historical linguistics – being central to models of language and cultural evolution – but also in downstream applications relying on cognates, such as machine translation.

COGNATES are words in sister languages (languages descending from a common ancestor) with a common proto-word. For example, the Romanian word *victorie* and the Italian word *vittoria* are cognates, as they both descend from the Latin word *victoria* 'victory' – see Figure 7.1. Cognates can help students when learning a second language and contribute to the expansion of their vocabularies. Cognate sets have also been used in a number of applications in natural language processing, including, for example, machine translation (Zhao & Zhang 2018). These applications rely on properly distinguishing between true cognates and false friends.

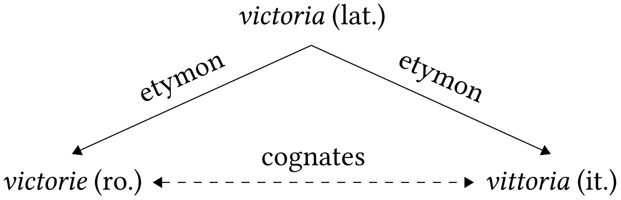


Figure 7.1: Example of cognates and their common ancestor.

In most cases, cognates have preserved similar meanings across languages, but there are also exceptions. In some cases, the meanings of cognates have diverged from the common etymon through their use in each of the two languages, and their meanings became different from each other. These are called *deceptive cognates* or, more commonly, *false friends*. Here we use the definition of cognates that refers to words with similar appearance and some common etymology, while *true cognates* is used to refer to cognates which also have a common meaning, and DECEPTIVE COGNATES or FALSE FRIENDS refers to cognate pairs which do not have the same meaning (anymore).

Many false friends have diverged into entirely different meanings. There are many examples, however, for which the changes in meaning are more subtle (for example, in connection with the feeling attached to a word or at the level of connotations) and more difficult to detect even for humans. The notion of semantic equivalence used to define false friends is in itself ambiguous and difficult to treat as a binary property, and we propose in this chapter that the quality of a cognate pair of being in a false friends relationship should also be treated as a spectrum. Based on this observation, we define the notions of HARD FALSE FRIEND and SOFT FALSE FRIEND.

HARD FALSE FRIENDS are cognates whose meanings have diverged enough such that they do not have the same sense anymore, and should not be used interchangeably (as translations of one another). Most known examples of false friends fall in this category, such as the French-English cognate pair *attendre*/*attend*: in French, *attendre* has a completely different meaning, which is 'to wait'. A different and more subtle type of false friends is represented by SOFT FALSE FRIENDS, which can result from minor semantic shifts between the cognates. In such pairs, the meaning of the words may remain roughly the same, but with a difference in nuance or connotation. Such an example is the Romanian-Italian cognate pair *amic*/*amico*. Here, both cognates mean 'friend', but in Italian the connotation is that of a closer friend, whereas the Romanian *amic* denotes a more distant friend, or even acquaintance. A more suitable Romanian translation for *amico* would be *prieten*, while a better translation in Italian for *amic* could be *conoscente*. Though their meaning is roughly the same, translating one word for the other would be an inaccurate use of the language. These cases are especially difficult to handle by beginner language learners (especially since the cognate pair may appear as a valid translation in multilingual dictionaries). In these cases, instead of helping non-natives to more easily understand a text in a foreign language, cognates can instead cause more confusion and deceive the language learner into misunderstanding the text, as using them in the wrong contexts is an easy trap to fall into.

Given these considerations, an automatic method for finding the appropriate term to translate a cognate into instead of using a false friend would be useful for assisting with language learning and text comprehension in a foreign language. Moreover, identifying false friends can be useful not only for language acquisition, but also in downstream applications relying on cognates, such as machine translation.

1.1 Related work

Cross-lingual semantic word similarity consists in identifying words that refer to similar semantic concepts and convey similar meanings across languages (Vulić & Moens 2013). Some of the most popular computational approaches developed for this task rely on probabilistic models (Vulić & Moens 2014) and cross-lingual word embeddings (Søgaard et al. 2017). In this area, a fundamental task is that of Bilingual Lexicon Induction (Mikolov et al. 2013, Heyman et al. 2017, Vulić & Moens 2015), which aims to discover new translations at the lexical level by automatically mapping between vector spaces of languages.

A comprehensive list of cognates and false friends for every language pair is difficult to find or manually build. Moreover, dictionaries grow outdated and it is difficult to continuously update them to incorporate new words in the vocabulary. This is why applications have to rely on automatically identifying false friends.

There have been a number of previous studies attempting to automatically extract pairs of true cognates and false friends from corpora or from dictionaries. Most methods are based either on orthographic and phonetic similarity, or require large parallel corpora or dictionaries (Inkpen et al. 2005, St Arnaud et al. 2017, Nakov et al. 2009, Chen & Skiena 2016).

Inkpen et al. (2005) use orthographic features to extract French-English cognate pairs, but do not take semantic similarity into account. Torres & Aluísio (2011) also rely on orthographic and phonetic features, to which they add a semantic feature extracted from a bilingual dictionary. They additionally release a lexicon of Spanish-Portuguese false friends and true cognates, obtained through manual annotation, that they use to evaluate their algorithms. Nakov et al. (2009) identify false friends pairs in Bulgarian and Russian by making use of sentence-aligned parallel corpora. Aminian et al. (2015) propose using a model of identifying false friends from parallel corpora in order to improve English-Egyptian statistical machine translation.

There have been few previous studies using word embeddings for the detection of false friends or cognate words, usually using simple methods on only one or two pairs of languages (Castro et al. 2018, Torres & Aluísio 2011). Castro et al. (2018) detect false friends in Spanish-Portuguese, employing a classifier that learns from features extracted from multilingual embedding spaces. Mitkov et al. (2007) use a method based on distributed representations of words in a continuous space built using comparable corpora, as well as a taxonomy-based approach, to identify false friends in four language pairs involving English, French, German and Spanish.

In recent years, multiple computational linguistic studies have focused on the issue of semantic change, tracking the shift in the meaning of words by looking at their usage across time in corpora dating from different time periods. More than this, computational linguists have also tried to systematically analyze the principles describing semantic change hypothesized by linguists (such as the law of parallel change and the law of differentiation; Xu & Kemp 2015), or even proposed new statistical laws of semantic change, based on empirical observations, such as the law of conformity (stating that polysemy is positively correlated with semantic change), the law of innovation (according to which word frequency is negatively correlated with semantic change; Hamilton et al. 2016), or the law of prototypicality (according to which prototypicality is negatively correlated with semantic change; Dubossarsky et al. 2015). More recently, Dubossarsky et al. (2017) revisited some of the semantic change laws proposed in previous literature, claiming that a more rigorous consideration of the control conditions when modelling these laws leads to the conclusion that they are weaker or less reliable than reported. More extensive surveys of computational studies relating to semantic change have been conducted by Kutuzov et al. (2018) and Tahmasebi et al. (2018) (see also Tahmasebi et al. 2021).

All previous computational studies on lexical semantic change have, to our knowledge, only looked at the semantic change of the words in monolingual settings, and where more than one language was included, they were considered independently (Hamilton et al. 2016). However, words do not evolve only in their own language in isolation, but are rather inherited and borrowed between and across languages. Dominguez & Nerlich (2002) distinguish between *chance false friends*, which have similar form but different etymologies as well as different meanings in different languages, and *semantic false friends*, which share the etymological origin, but their meanings differ (to some extent) in different languages. In this study we focus on the latter, which we consider more relevant from the point of view of semantic change since, in principle, they begin with a common meaning then diverge, to a lower or higher degree, while often preserving some common meaning, whereas *chance false friends* usually have entirely distinct meanings.

Uban et al. (2019a) propose a method for identifying and correcting false friends and define a measure of their "falseness", using cross-lingual word embeddings. We base our study on the method proposed there and take it further by analyzing the properties of semantic divergence as they relate to different properties of the words, across five Romance languages, as well as English. Similarly to how Hamilton et al. (2016) formulate statistical laws of semantic change within one language, describing how word properties affect semantic change, we propose studying analogous laws cross-lingually, from the point of view of cognate divergence, studying this time the effect of semantic divergence on the subsequent evolution of the words, including properties such as frequency and polysemy. When a word enters a new language, features specific to that particular language (such as existing words in the same language or socio-cultural and historical factors) can affect the way it is used and contribute to shaping its meaning through time. The evolution of cognate words in different languages can be seen as a collection of different parallel histories of the proto-word from its entering the new languages to its current state. Based on this view, we propose a novel approach for studying semantic change: instead of comparing *monolingual* texts from *different time periods* as ways to track meanings of words at different stages in time, we compare *present meanings* of cognate words across *different languages*, viewing them as snapshots in time of each of the word's different histories of evolution. We expand upon the work published in Uban et al. (2019b) and continue exploring how properties (namely frequency and polysemy) of the words involved in semantic change relate to the degree of their semantic shift, and find the concrete mathematical functions that best describe the relationship. Additionally, we present examples of true cognates for each language (words which kept their Latin meaning in the modern language), and cluster them into semantic fields, which might provide some insight into the socio-cultural factors that are connected to semantic change.

1.2 Contributions

The contributions of our work on cognate divergence are threefold. Firstly, we propose a method for quantifying the semantic divergence of languages. Secondly, we provide a framework for detecting and correcting false friends, based on the observation that these are usually deceptive cognate pairs: pairs of words that once had a common meaning, but whose meaning has since diverged. Thirdly, we propose a novel way to measure semantic change synchronically across languages, by tracking the divergence of cognate words from their original etymon.

In Section 2, we introduce a method for measuring the semantic divergence of sister languages based on cross-lingual word embeddings. We use a multilingual set of cognates extracted from etymology dictionaries and word embeddings trained on Wikipedia corpora. By comparing current meanings of cognate sets in different languages, our method can uncover insights about how their meanings diverged within their respective languages from their common original etymon, and infer properties of the parallel processes of change in the meaning of cognate words across time. We report empirical results on five Romance languages: Romanian, French, Italian, Spanish and Portuguese. For a deeper insight into the matter, we also compute and investigate the semantic similarity between modern Romance languages and Latin. We then introduce English into the mix, to analyze the behavior of a more remote language, where words deriving from Latin are mostly borrowings. We show that, in terms of semantic divergence, the studied languages form clusters that are consistent with the generally accepted tree of languages. Moreover, we perform a qualitative analysis of the subset of cognates for which meaning was preserved from the etymon to the modern word, comparatively between different language pairs, and show how the original Latin meaning of words was preserved across Romance languages for different semantic fields. In Section 3, we propose a fully automated, unsupervised method for false friend detection and correction, relying on cross-lingual word embeddings. We propose a corpus-based approach that is capable of covering the majority of the vocabulary for a large number of languages, while at the same time requiring minimal human effort in terms of manually evaluating word pair similarity or building lexicons, relying only on large monolingual corpora. We propose a method that can be used to identify pairs of false friends, to distinguish between the two categories of false friends defined above (*hard false friends* and *soft false friends*), and to provide suggestions for correcting the erroneous usage of a false friend in translation. We evaluate the algorithm on Romance languages and English. We build a dataset of false friends, publicly available, along with falseness scores for each pair. In Section 4, we propose a method for measuring and characterizing semantic change using the semantic divergence of cognate sets. Building on related literature in computational linguistics, we study how laws of semantic change manifest cross-linguistically, trying to understand how semantic divergence affects word properties in the multilingual setting, from a reversed perspective compared to previous studies: namely measuring the effect of semantic change on word properties (such as frequency and polysemy). We show that, from this perspective, semantic divergence is positively correlated with both polysemy and frequency. In Section 5, we draw conclusions and discuss future work.

1.3 Cross-lingual word embeddings

Word embeddings are vectorial representations of words in a continuous space, built by training a model to predict the occurrence of a word in a text corpus, given its context, or the context, given the word. Based on the distributional hypothesis stating that similar words occur in similar contexts, these vectorial representations can be seen as semantic representations of words and can be used to compute semantic similarity between word pairs (representations of words with similar meanings are expected to be close together in the embedding space).

To compute the semantic divergence of cognates across sister languages, as well as to identify pairs of false cognates (pairs of cognates with high semantic distance), which by definition are pairs of words in two different languages, we need to obtain a multilingual semantic space, which is shared between the cognates. Having the representations of both cognates in the same semantic space, we can then compute the semantic distance between them using their vectorial representations in this space. Our research is related, in terms of methodology, to Bilingual Lexicon Induction, which has been extensively studied in previous research (Mikolov et al. 2013, Heyman et al. 2017), but in our case, we rely on inferred cross-lingual lexical semantic similarities in order to verify whether cognate pairs share the same meaning, rather than discover new translations.

For our purposes, we use the publicly available FastText (Bojanowski et al. 2017) multilingual word embeddings, pre-trained on Wikipedia for the six languages in question, and pre-aligned in a common vector space (Conneau et al. 2017).[1] The vectors have 300 dimensions and were obtained using the skip-gram model described by Bojanowski et al. (2017) with default parameters.

The algorithm for measuring the semantic distance between cognates in a pair of languages (*lang1*, *lang2*) consists of the following steps:

1. Obtain word embeddings for each of the two languages.

2. Obtain a shared embedding space, common to the two languages. This is accomplished using an alignment algorithm, which consists of finding a linear transformation between the two spaces, that on average optimally transforms each vector in one embedding space into a vector in the second embedding space, minimizing the distance between a few seed word pairs (for which it is known that they have the same meaning), based on a small bilingual dictionary. For our purposes, we use the publicly available multilingual alignment matrices that were published by Smith et al. (2017).

3. Compute semantic distances for each pair of cognate words in the two languages, using a vectorial distance (we chose cosine distance) on their corresponding vectors in the shared embedding space.

When interpreting results based on aligned embedding spaces to infer conclusions on linguistic phenomena at the language level, various limitations of

[1] https://github.com/facebookresearch/MUSE

the method should be kept in mind, including biases due to the corpus used to train the embeddings, to the algorithm used to train the embeddings, or to the alignment operation. Unsupervised alignment of pre-trained monolingual embedding spaces for obtaining multilingual representations has been shown in previous studies to introduce noise in the resulting multilingual space (Søgaard et al. 2018, Beinborn & Choenni 2019, Patra et al. 2019), due to the simplifying assumptions on the isomorphy of monolingual embedding spaces, and to properties of the monolingual spaces themselves, leading to different alignment quality for different language pairs. We attempt to minimize the effect of the confounding factors through our particular methodological choices, where possible, and through experiments designed to measure the contribution of the different factors in isolation.

We choose the FastText (Bojanowski et al. 2017) embeddings pre-trained on Wikipedia since they are trained on large amounts of text, which minimizes the amount of noise in the vectors, making them good approximators of word meanings. Additionally, they are trained on text that is relatively uniform in style and topic, which ensures that any difference in the structure of the embedding spaces of different languages depends on the language, rather than being an artifact of topic or genre. Nevertheless, even high quality embeddings can be noisy or biased and this should be kept in mind when interpreting the results of our experiments. Moreover, the uniformity of the writings in the corpus in terms of style and period of history when they were written can act as a weakness, limiting the embeddings' representativeness of the language as a whole (Koplenig 2016). For Latin in particular, we note that using Latin Wikipedia as a training corpus might bias representations away from the original usages of the words in Latin, and towards their usages in the modern languages the articles were translated from.

The algorithm that we use for computing semantic distance for cognate pairs stands on the assumption that the (shared) embedding spaces are comparable, so that the averaged cosine similarities and the overall distributions of scores that we obtain for each pair of languages can be compared in a meaningful way. For this to be true, at least two conditions need to hold:

1. The embedding spaces for each language need to be similarly representative of the language, or trained on similar texts – this assumption holds sufficiently in our case, since all embeddings (for all languages) are trained on Wikipedia, which at least contains a similar selection of texts for each language, and at most can be considered comparable corpora.

2. The similarity scores in a certain (shared) embedding space need to be sampled from a similar distribution. To confirm this assumption, we compare distributions of a random sample of similarity scores across all embedding spaces. For each multilingual embedding space (corresponding to a language pair), we select at random 1,000,000 word pairs, and compute their similarities. We find that the similarity distributions are similar in mean and standard deviation across aligned embedding spaces, with mean similarity scores ranging between $(0.188, 0.199)$ for all language pairs, and all standard deviations between $(0.074, 0.082)$. On our (large) samples of word pairs used in this analysis, statistical t-tests show significant ($p < 0.05$) difference between means across language pairs, but the effect is small: Cohen's d-test shows an effect size smaller than 0.09 for all language pairs.

   The observed consistency of word similarities across language pairs was not obvious but also not surprising, since:

   - The way we create shared embedding spaces is by aligning the embedding space of any language to the English embedding space (which is a common reference to all shared embedding spaces).
   - The nature of the alignment operation (consisting only of rotations and reflections) guarantees monolingual invariance, as described by Artetxe et al. (2016) and Smith et al. (2017).

2 The semantic divergence of cognates

We propose a definition of semantic divergence between two languages based on the semantic distances of their cognate word pairs in embedding spaces. The semantic distance between two languages can then be computed as the average semantic divergence of all pairs of cognates in that language pair.

As our data source for cognate words, we use the list of cognate sets in Romance languages proposed by Ciobanu & Dinu (2014). It contains 3,218 complete cognate sets in Romanian, French, Italian, Spanish and Portuguese, along with their Latin common ancestors. The cognate sets are obtained from electronic dictionaries which provide information about the etymology of the words. Two words are considered cognates if they have the same etymon (i.e., if they descend from the same word). A subset of 305 of these sets also contains the corresponding cognate (in the broad sense, since these are mostly borrowings) in English.

One complete example of a cognate set for the word *architect* in the Romance languages is illustrated in Table 7.1.

Table 7.1: An example of a cognate set: *architect* in Romance languages.

Romanian	French	Italian	Spanish	Portuguese	Latin ancestor
arhitect	architecte	architetto	arquitecto	arquiteto	architectus

2.1 The Romance languages

We compute the cosine similarity between cognates for each pair of modern languages, and between modern languages and Latin as well. We compute an overall score of similarity for a pair of languages as the average similarity for the entire dataset of cognates. The results are reported in Table 7.2.

Table 7.2: Average cross-lingual similarity between cognates (Romance languages).

	Fr	It	Pt	Ro	La
Es	0.67	0.69	0.70	0.58	0.41
Fr		0.66	0.64	0.56	0.40
It			0.66	0.57	0.41
Pt				0.57	0.41
Ro					0.40

We observe that the highest similarity is obtained between Spanish and Portuguese (0.70), while the lowest values are obtained for Latin. From the modern languages, Romanian has, overall, the lowest degrees of similarity to the other Romance languages. A possible explanation for this result is the fact that Romanian developed far from the Romance kernel, being surrounded by Slavic languages. In Table 7.3 (page 231) we report, for each pair of languages, the most similar (above the main diagonal) and the most dissimilar (below the main diagonal) cognate pair for Romance languages.

The problem that we address in this experiment involves a certain *vagueness of reported values* (also noted by Eger et al. 2016 in the problem of semantic language classification), as there is no gold standard that we can compare our results to. To overcome this drawback, we use the degrees of similarity that we obtained to produce a language clustering, using the UNWEIGHTED PAIR GROUP METHOD WITH ARITHMETIC MEAN (UPGMA) hierarchical clustering algorithm (Sokal &

Michener 1958). We observe that it is similar to the generally accepted tree of languages, and to the clustering tree built on intelligibility degrees by Dinu & Ciobanu (2014). The obtained dendrogram is rendered in Figure 7.2.

2.2 The Romance languages vs. English

In this subsection, we introduce English into the mix. We run this experiment on a subset of the dataset of cognates, using only the words that have a cognate in English as well.[2] The subset has 305 complete cognate sets.

The results are reported in Table 7.4, and the distribution of similarity scores for each pair of languages is rendered in Figure 7.3. We notice that English has a comparatively low similarity with Latin (0.40 similarity with Latin, the same as French and Romanian), but its cognates are close to the other languages. Out of the modern Romance languages, Romanian is the most distant from English, with 0.53 similarity.

Another interesting observation relates to the distributions of scores for each language pair, shown in the histograms in Figure 7.3. While similarity scores between cognates among Romance languages usually follow a normal distribution (or another unimodal, more skewed distribution), the distributions of scores for Romance languages with English seem to follow a bimodal distribution, pointing to a different semantic evolution for words in English that share a common etymology with a word in a Romance language. One possible explanation is that the set of cognates between English and Romance languages (which are pairs of languages that are more distantly related) consist of two distinct groups: words that were borrowed directly from the Romance language to English (which should have more meaning in common), and words that had a more complicated etymological trail between languages (and for which meaning might have diverged more, leading to lower similarity scores).

Beinborn & Choenni (2019) have shown that when comparing a list of core words between languages in aligned embedding spaces, the average similarities differ between different language pairs, due to artifacts of the aligned embedding space itself, which is a possible confounding factor for our results. Translation pairs in two languages tend to be closer in the embedding space for more similar languages. Thus, any observed difference between average cognate similarity scores across language pairs could also be explained by the underlying properties of the embedding spaces, at the level of the overall vocabulary, and not an effect of semantic divergence between cognate pairs specifically.

[2] Here we "stretch" the definition of *cognates*, as they are generally referring to sister languages. In this case English is not a sister of the Romance languages, and the words with Latin ancestors that entered English are mostly borrowings.

Table 7.3: Most similar and most dissimilar cognates for all language pairs.

	Es	Fr	It	Ro	Pt
Es	–	ocho/ huit (0.89)	diez/ dieci (0.86)	ocho/ opt (0.82)	ocho/ oito (0.89)
Fr	caisse/ casar (0.05)	–	dix/ dieci (0.86)	décembre/ decembrie (0.83)	huit/ oito (0.88)
It	prezzo/ prez (0.06)	punto/ ponte (0.09)	–	convincere/ convinge (0.75)	convincere/ convencer (0.88)
Ro	miere/ mel (0.09)	face/ facteur (0.10)	as/ asso (0.11)	–	opt/ oito (0.83)
Pt	prez/ preço (0.05)	pena/ paner (0.09)	preda/ prea (0.08)	linho/ in (0.05)	–

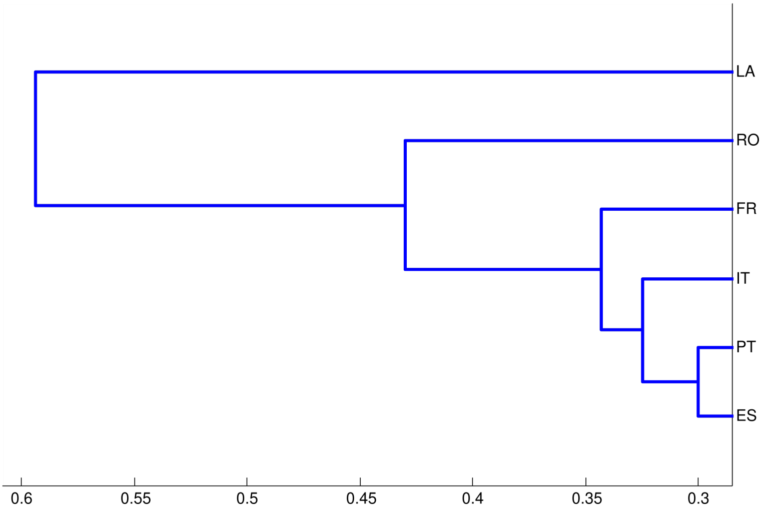
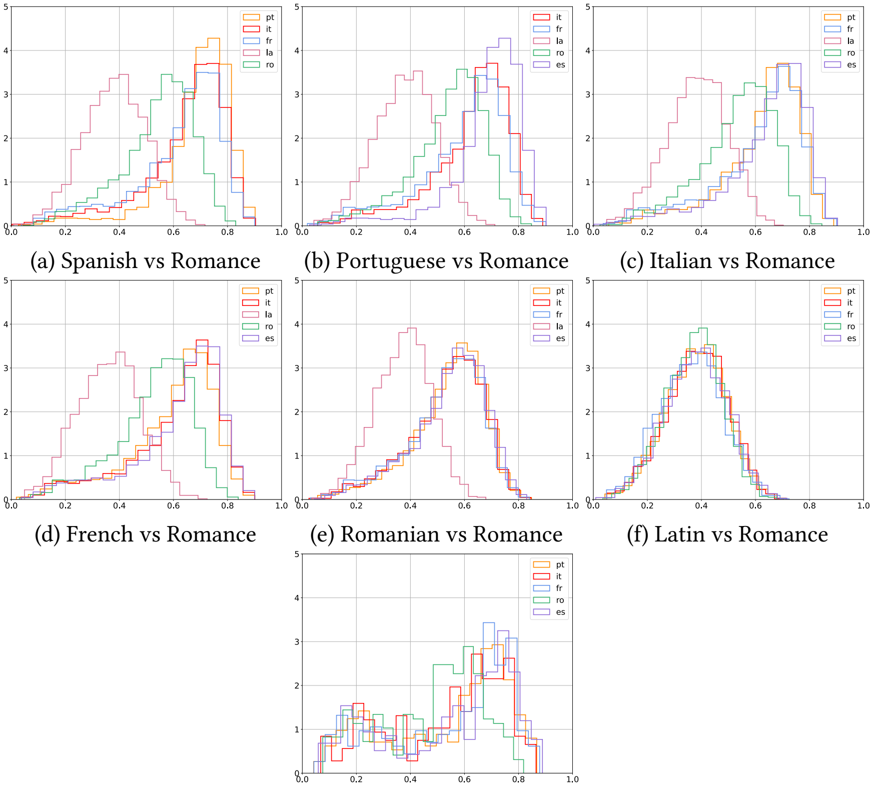


Figure 7.2: Dendrogram of the language clusters.

Table 7.4: Average cross-lingual similarity between cognates.

	Fr	It	Pt	Ro	En	La
Es	0.64	0.67	0.68	0.57	0.61	0.42
Fr		0.64	0.61	0.55	0.60	0.40
It			0.65	0.57	0.60	0.41
Pt				0.56	0.59	0.42
Ro					0.53	0.40
En						0.40



(a) Spanish vs Romance

(b) Portuguese vs Romance

(c) Italian vs Romance

(d) French vs Romance

(e) Romanian vs Romance

(f) Latin vs Romance

(g) English vs all

Figure 7.3: Distributions of cross-lingual similarity scores between cognates.

We further attempt to test the validity of our hypothesis that the observed cognate-based similarities between languages are, at least to some degree, specific to cognates, and thus can be interpreted as reflecting the process of cognate divergence. We compare the obtained similarities based on cognate pairs with baseline similarities between manually built translation pairs, using the seed dictionaries in Conneau et al. (2017), which contain approximately 100,000 word pairs for each language pair (between 87,000 and 113,00 across all included language pairs). We compute these for the language pairs where seed dictionaries were available (so excluding Romanian and Latin). A difference between similarities based on translation pairs and similarities based on cognate pairs would confirm that cognate divergence does contribute the observed effect (which is not simply due to embedding space alignment). The baseline used here differs from the one reported in Section 1 where we sample word similarities across the entire embedding space randomly, as opposed to focusing on translation pairs.

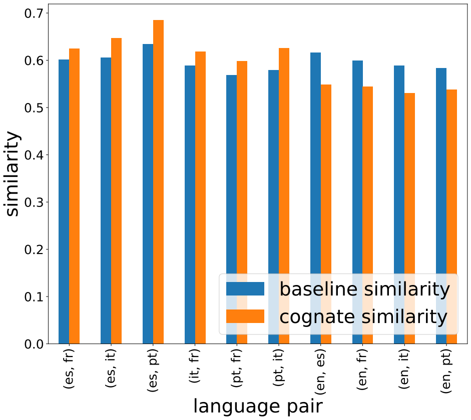


Figure 7.4: Comparison of language similarity scores based on baseline dictionaries and on cognate pairs.

Table 7.5 shows average similarities for seed translation pairs, and Figure 7.4 illustrates direct comparisons of average similarities for baseline terms and for cognate pairs. Here we notice a pattern of higher cognate-based similarity for Romance languages compared to the baseline, and lower cognate similarities for pairs involving English.

Beyond these differences, we notice however that translation similarities are not uniform across language pairs either, showing a similar pattern to cognate similarities. This would entail that the observed effect in differences between language pairs, and hypothesized higher similarity between closer languages (such as Spanish and Portuguese), is at least partly due to imperfect embeddings alignment noise. To compare the contribution of embedding alignment noise to that of cognate divergence, we compute the Spearman correlation between the cognate-based distances and the translation-based distances for every language pair (for a total of 10 language pairs), in order to verify whether the relative order of language pairs in terms of average similarities is different in cognates and seed translations, and we obtain a correlation of 0.36 (p-value 0.32). We conclude that there is a noticeable effect related to cognate divergence, but that alignment noise should be kept in mind when interpreting the results, and further research based on more refined alignment algorithms or a different class of methods for measuring cross-lingual semantic distance would be useful for a more conclusive result. We leave for future work a similar analysis for the missing language pairs, including Romanian and Latin, with the possible inclusion of additional baseline word pairs, such as the ones used in Beinborn & Choenni (2019).

Table 7.5: Average cross-lingual similarity between seed translation pairs.

	Fr	It	Pt	En
Es	0.60	0.60	0.63	0.61
Fr		0.58	0.56	0.59
It			0.57	0.58
Pt				0.58

3 Detection and correction of false friends

In this section, we propose using a fully automatic and unsupervised algorithm in order to detect false friends, and we generate a lexicon of false friends, along with falseness scores for each pair, for every language pair among six considered languages (five Romance languages and English). Our method is based on the false friend detection algorithm relying on cross-lingual word embeddings introduced by Uban et al. (2019a), to which we add a more extensive evaluation of the resulted false friends pairs, including the extended list of over 3,000 cognate sets (instead of the smaller 305 words list evaluated in the previous study) and additionally include an evaluation and analysis of the falseness measure. We publish freely the resulting database comprising of false friend pairs for each pair of considered languages, and the falseness score for each pair.

Our chosen method of leveraging word embeddings extends naturally to another application related to this task which, to our knowledge, has not been explored so far in research: false friend correction. We propose a straightforward method for solving this task of automatically suggesting a replacement when a false friend is incorrectly used in a translation. Solving this problem could result in a tool especially useful for language learners to help them use language correctly.

3.1 Algorithm

In the following subsection, we describe the algorithm used for detecting false friends automatically, in an unsupervised manner, based on a seed set of cognate sets, as well as a method for correcting false friends. Using the same principles as in the previous experiment, we can use embedding spaces and semantic distances between cognates in order to detect pairs of false friends, which are simply defined as pairs of cognates which do not share the same meaning, or which are not semantically similar *enough*.

False friends can be identified as pairs of cognates with high semantic distance. More specifically, we consider a pair of cognates to be a false friend pair if in the shared semantic space, there exists a word in the second language which is semantically closer to the original word than its cognate pair in that language (in other words, the cognate is not the optimal translation). The arithmetic difference between the semantic distance between these words and the semantic distance between the cognates will be used as a measure of the *falseness* of the false friend.

The algorithm has the additional ability to provide suggestions for correcting false friends: the nearest neighbor (in the second language) to the first cognate will be the suggested "correction", which should correspond to the correct translation of the cognate. This solution is based on the principle used in BILINGUAL LEXICON INDUCTION (BLI), which is the task of automatically discovering words with the same meaning across languages (Mikolov et al. 2013, Heyman et al. 2017): our algorithm can be seen as an application of BLI. In previous literature, multilingual embedding spaces have been used for BLI (Vulić & Moens 2015); we propose they can be useful in the context of false friends and language learning. The approach is described in detail in Algorithm 1.

We select a few results of the algorithm to show in Table 7.6, containing examples of extracted false friends, along with the suggested correction and the computed degree of falseness. The table shows some examples of the algorithm correctly identifying and correcting false friends pairs – such as the Romanian-Italian pairs *tânăr* 'young'/*tenero* 'tender', with the Italian correction *giovane* 'young', or *inimă* 'heart'/*anima* 'soul', corrected to *cuore* 'heart'. The falseness

Algorithm 1: Detection and correction of false friends

```
Given the cognate pair $(c_1, c_2)$ where $c_1$ is a word in $lang_1$ and $c_2$ is a word in $lang_2$:
Find the nearest neighbor of $c_1$ in $lang_2$ as the word $w_2$ in $lang_2$ such that for any $w_i$ in $lang_2$, $distance(c_1, w_2) < distance(c_1, w_i)$
if $w_2 \neq c_2$ then
    $(c_1, c_2)$ is a pair of false friends
    Degree of falseness = $distance(c_1, w_2) - distance(c_1, c_2)$
    return $w_2$ as potential correction
end
```

Table 7.6: Extracted false friends and falseness.

Cognate	False friend	Correction	Falseness
long (Fr)	luengo (Es)	largo	0.50
face (Fr)	faz (Es)	cara	0.39
change(Fr)	caer (Es)	cambia	0.46
stânga (Ro)	stanco (It)	destra	0.52
tânăr (Ro)	tenero (It)	giovane	0.41
inimă (Ro)	anima (It)	cuore	0.13
amic (Ro)	amico (It)	amichetto	0.04

scores also reflect the degree of semantic drift between the false friends, with the *tânăr*/*tenero* pair being more dissimilar than *inimă*/*anima*. The *amic*/*amico*/*amichetto* set, which refers to different degrees of friendship, is awarded the lowest falseness score. It is valuable to note the algorithm also selects word pairs which can technically be considered true cognates (*long*/*luengo* – meaning 'long', but are not used as such in current speech: *largo* is more frequently used than *luengo*. This is to be expected since the algorithm is based on word *usage* in language (since this is the basis of the embedding training algorithm). We also illustrate an example where the algorithm makes a mistake: in the case of *stânga* 'left'/*stanco* 'tired', the algorithm rightly identifies this as a false friends pair, but provides an erroneous correction: *destra* is the Italian word for 'right', not 'left'. This error can also be traced back to the nature of semantic similarity as captured by word embeddings: related but not equivalent words (and sometimes even antonyms) can have similar embedding vectors due to their similar occurrence patterns in corpora.

3.2 Building a false friends dataset

We use the algorithm described in the previous subsection to build a database of false friends pairs for each language pair among the six considered languages, which we make freely available.[3] False friends for Romance languages are extracted from the original 3,218 cognate sets, resulting in 500 to 1,200 pairs of detected false friends for each language pair. For English, the original cognate resource contains a smaller set of only 305 cognate sets, which results in smaller false friends lists for language pairs involving English. Table 7.7 shows the number of false friends pairs generated for each language pair, and included in the published resource.

Table 7.7: Number of datapoints in false friends database.

Languages	FF Pairs	Languages	FF Pairs
Es–It	739	It–Es	727
Es–Pt	490	Pt–Es	502
Fr–It	921	It–Fr	925
Fr–Es	886	Es–Fr	905
Fr–Pt	1,023	Pt–Fr	1,060
It–Pt	795	Pt–It	848
Ro–Fr	1,258	Fr–Ro	1,596
Ro–It	1,286	It–Ro	1,654
Ro–Es	1,229	Es–Ro	1,647
Ro–Pt	1,227	Pt–Ro	1,640
En–Pt	148	Pt–En	137
En–Es	158	Es–En	136
En–It	153	It–En	139
En–Fr	150	Fr–En	133
En–Ro	205	Ro–En	161

3.3 Evaluation

In order to evaluate the quality of the false friends dataset generated with our algorithm, we first test its accuracy against a multilingual dictionary. For this study, we choose to use Open Multilingual WordNet (Miller 1998, Bond & Foster

[3] https://github.com/ananana/false_friends_resource

2013). WordNet is a semantic network organized in synsets which represent concepts, where each word is part of as many synsets as concepts it designates. Two words with common etymology are considered true cognates if they belong to the same WordNet synset (are synonyms), and false friends if they are found in WordNet, but not as synonyms. Cognates not found in any WordNet synset are not considered. Using this standard, the obtained measured accuracy is between 73% and 81%, depending on the considered language pair. Table 7.8 presents a breakdown of the obtained performance per considered language pair. Romanian is the only language missing from the evaluation since it is not represented in multilingual WordNet. Since English cognates are only available for a subset of the cognates list, our evaluation results for Romance languages may be more robust.

Table 7.8: Performance for all language pairs using WordNet as the gold standard.

	Accuracy	Precision	Recall
Es–It	73.69	43.27	38.06
It–Es	73.58	43,12	37.73
Es–Pt	79.09	36.05	26.49
Pt–Es	78.65	32.32	24.35
Fr–It	74.43	33.39	57.40
It–Fr	74.77	34.32	58.68
Fr–Es	76.25	42.02	51.94
Es–Fr	75.13	40.27	51.78
It–Pt	74.58	33.20	44.73
Pt–It	73.61	31.69	49.31
En–Pt	77.25	59.81	86.48
Pt–En	79.82	64.70	85.71
En–Es	76.58	63.88	88.46
Es–En	80.48	71.57	83.95
En–It	77.40	61.73	87.65
It–En	74.89	61.90	76.47
En–Fr	77.09	57.89	94.28
Fr–En	81.05	66.32	86.66

In a second experiment, we measure the accuracy of false friend detection on a manually curated list of false friends and true cognates in Spanish and Portuguese, used in a previous study (Castro et al. 2018), and introduced by Torres & Aluísio (2011). This resource is composed of 710 Spanish-Portuguese word pairs: 338 true cognates and 372 false friends. We also compare our results to the ones reported in this study, which uses a method similar to ours (using a simple classifier that takes embedding similarities as features to identify false friends) and shows improvements over results in previous research. The results are shown in Table 7.9. We also compute the same metrics using a falseness threshold as a lower bound to decide whether two words are false friends, and observe a trade-off between recall and precision when using a threshold. The following subsection discusses the use of falseness thresholds in more detail.

In this second experiment, WordNet is used as a baseline algorithm for false friend identification instead of a gold standard. Its relatively poor results (reported in Table 7.8), in comparison with the automatic methods, may stem from its coverage, which is lower than for corpus-based methods. Castro et al. (2018) show that only 55% of the word pairs in the evaluation set used here are found in WordNet synsets. This shows that using WordNet as an evaluation standard has its limits, and that corpus-based methods for evaluating cross-lingual semantic similarity, such as the one we propose, have an advantage over dictionary-based methods.

Table 7.9: Performance for Spanish-Portuguese using curated false friends test set, compared to previous attempts.

	Accuracy	Precision	Recall
Our method (ft = 0)	81.81	78.69	80.80
Our method (ft = 0.1)	82.62	92.37	66.06
Castro et al. (2018)	77.28	–	–
Torres & Aluísio (2011)	76.37	–	–
WN Baseline	69.57	85.82	54.50

3.3.1 Falseness as a spectrum

The measure of falseness that we provide for every detected pair of false friends can be useful not only for a better understanding of the linguistic phenomenon behind the semantic divergence of the cognates, but also for a more flexible integration with downstream applications. When using our resource of false friends, a custom threshold of falseness could be set for filtering out false friends in a more coarse or fine-grained way, depending on the needs of the application: by selecting as false friends only pairs with falseness above a specific (non-zero) threshold. For example, for applications where capturing subtle changes in meaning is important, maintaining a low threshold of falseness is useful. On the other hand, when the goal is to only identify false friends which have entirely different meanings, choosing a high threshold may be sufficient. It might also ensure a lower rate of false positives by filtering out the delicate cases of cognates which lie at the boundary between true and deceptive cognates.

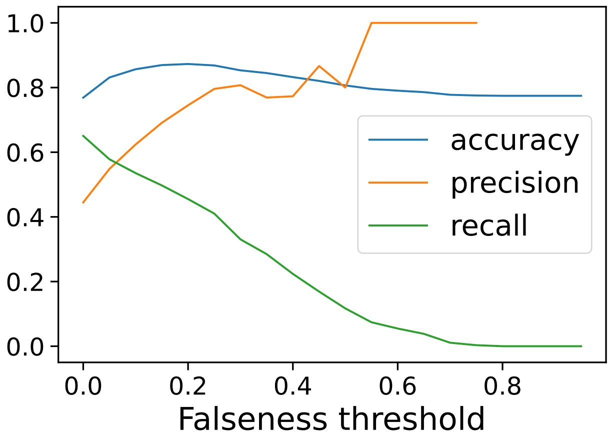


Figure 7.5: Performance with falseness threshold.

We perform an analysis of the effect of varying the threshold applied to the falseness score in order to discriminate between true cognates and false friends, by re-evaluating the generated false friends against WordNet using different threshold values (as opposed to the simple evaluation in the previous section, where no threshold was set, which is equivalent to using a falseness threshold of 0). In this way, we are able to discover the optimal falseness threshold to use in order to maximize performance relative to the WordNet standard. We choose the threshold which leads to maximum average accuracy across all language pairs on a separate training set of 80% of the word pairs. The rest of 20% of the word pairs are used to evaluate the method, now employing a falseness threshold set to the optimal value according to the training phase. The optimal falseness threshold $ft\star$ is found to be 0.2, and the average overall accuracy with this threshold is 85.85%. Table 7.10 shows the difference in accuracy when using the optimal threshold, and Figure 7.5 illustrates the variation of accuracy, precision and recall (on average for all language pairs) when varying the threshold between 0 and 1.

$$ft\star = \text{argmax}_{ft\in(0,1)} \frac{1}{|\text{LP}|} \sum_{l_1,l_2\in\text{LP}} \text{Acc}(ft, l_1, l_2) \tag{7.1}$$

where *ft* is a falseness threshold and LP is the set of all language pairs:

$$\text{LP} = \{(l_1, l_2)|l_1, l_2 \in \{\text{Ro, Es, Pt, Fr, It, En}\}\} \tag{7.2}$$

The fact that in WordNet the optimal falseness threshold is positive (non-zero) suggests that many of the pairs with very low falseness make up for most of false positives (actual true cognates) and are responsible for a drop in accuracy (as they are probably identified as false friends by the algorithm not necessarily because they are actually different in meaning, but rather because of artifacts of the embedding space).

Table 7.10: Best overall accuracy of our method.

Falseness threshold	0 (None)	0.2 (optimal)
Accuracy	80.57	85.85

We then perform the same experiment, but this time evaluate using the curated cognate sets in Spanish-Portuguese. In this case, the optimal threshold is found to be 0.1, and the threshold of 0.2 found in the previous experiment leads to worse results than not using a threshold at all. In order to confirm that the difference stems from the the different definition of cross-lingual synonymy in the two datasets and is not specific to just the language pair, we compute the optimal falseness threshold relative to WordNet specifically for Spanish-Portuguese and find an optimal value of 0.3. One explanation for the different optimal thresholds on the two reference datasets may be that they cover different parts of the vocabulary, as confirmed by the previously reported low coverage of WordNet on the Spanish-Portuguese test set. The difference between the optimal threshold values for the two different gold standards may also suggest that the two resources were built based on different assumptions about meaning equivalence, and confirms that the availability of the falseness measure can be useful for tuning the false friend detection algorithm to the specific task and standards of the particular application.

3.3.2 Error analysis and discussion

As suggested in the previous subsection, a significant source of error relative to the WordNet standard are low-falseness pairs of detected false friends. Figures 7.6 and 7.7 show the distribution of falseness scores across all word pairs in all languages. We separately show the distribution of false friends extracted with our method that were evaluated as actual false friends using WordNet, and the pairs of extracted false friends that are actually true cognates according to WordNet. The much lower falseness values for word pairs in the second category (false positives in the evaluation using WordNet) suggest that many of the false positives produced by the algorithm fall in the range of word pairs with very subtle differences in meaning. These might stem from imperfections in the embedding space or from the too strong assumption that the closest word in the multilingual embedding space is the correct translation. Some of the examples in Table 7.6 illustrate these types of error; such is the case of the previously discussed pair *stânga*/*stanco*, with the mistaken correction *destra*. More subtle inaccuracies can consist, for example, of mismatched parts of speech, such as the case of *change* (noun)/*caer* (infinitive verb)/*cambia* (indicative verb).

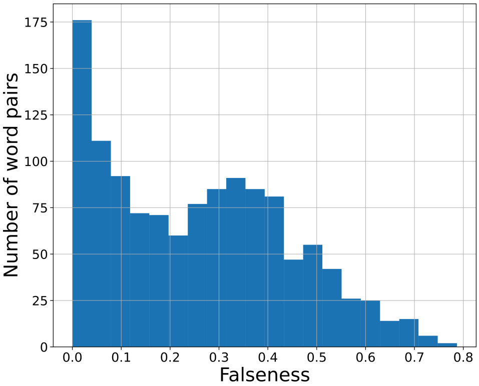


Figure 7.6: Falseness in correctly detected false friends.

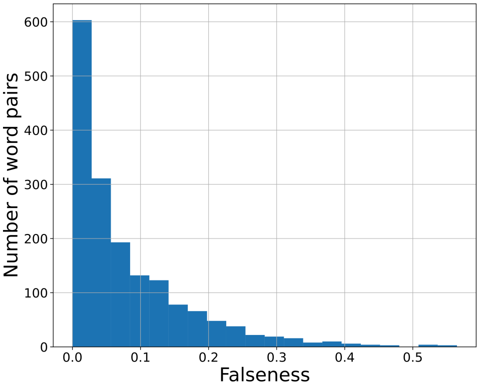


Figure 7.7: Falseness in incorrectly detected false friends.

We argue that including the falseness score in our published lexicon of false friends can be useful precisely to remedy this issue when needed, by setting a higher threshold on the falseness score. On the other hand, it is possible that in some cases, the low-falseness word pairs classified as false positives according to WordNet could even be considered actual false friends (rather than errors of classification) by a standard of meaning equivalence that is more strict than the one used in WordNet's synsets, again confirming the value of modelling falseness as a spectrum.

A second source of errors is found in the original cognates data source that we use to discriminate into true and false cognates. Since it is also an automatically built resource, some of the word pairs are falsely labelled as cognates, and may further perpetuate into false positives in our algorithm.

4 Laws of cross-lingual semantic change

We use the measure of falseness of a deceptive cognate pair to quantify the semantic shift between the meanings of a word derived from the same etymon in different languages. We further propose analyzing how the properties of frequency and polysemy of a word relate to semantic shift, and, analogously to what Hamilton et al. (2016) do for monolingual semantic change, we aim to move towards uncovering statistical laws of semantic change across languages.

In the next subsection, we first define a measure of the frequency of a word, as well as a measure of its polysemy. Further, we try to correlate these measures of frequency and polysemy with the falseness measure defined in the previous subsections. Finally, we find mathematical equations that best describe the relationship between frequency and polysemy of words in a cognate pair on one hand, and the falseness degree of the pair on the other hand, according to our dataset. As a preliminary step, we discard all cognate pairs that, according to the false friend detection algorithm, are true cognates, and focus only on the deceptive cognates, for which falseness scores are non-zero. On average across all language pairs, 37% of the cognate pairs in our dataset are found as deceptive cognates. Moreover, we validate these results using multilingual WordNet, and further select only pairs which are confirmed to be deceptive cognates as such: two cognates are considered to be true cognates if they are synonyms according to WordNet, and are considered to be deceptive cognates otherwise. It should be noted that having to use WordNet limits us to languages for which WordNet is available (which excludes Romanian, for which we consider all words in the cognate set instead).

Through characterizing the relationship between frequency and polysemy on the one hand, and semantic change (as measured by falseness in our case) on the other hand, we aim to discover statistical laws that describe how semantic change of words relates to other properties of the words. Similar attempts at formulating laws of semantic change have been made in previous studies in monolingual diachronic settings, with the notable example of Hamilton et al. (2016), who find polynomial relationships between the same word properties (frequency, polysemy) and the degree of semantic change over time. Nevertheless, an important difference is that while the authors of the monolingual study correlate the rate of the shift of meaning for a word to its frequency and polysemy *prior* to the change in meaning, our method looks at the magnitude of the meaning shift in comparison with properties of words *after* the meaning shift has already occurred, presumably from the original meaning of the proto-word they derive from to their current meanings in their respective languages.

4.1 Word frequency and semantic divergence

For measuring *frequency*, we use the multilingual Wordfreq Python library (Speer et al. 2018), which estimates word frequency based on multiple corpora (such as Wikipedia and Twitter). For most of the languages we consider, we are able to extract frequency scores for the majority of words in our cognate sets, with a coverage of at least 92% of the words in our cognate set for every language considered, except for Romanian, which has a poorer coverage of only 60%.

For each pair of languages in a cognate set, we compute the Spearman correlation between the average of the frequencies of the words in the cognate pair and the falseness of the deceptive cognate. Since frequency and polysemy are correlated, we need to control for polysemy in order to observe the marginal effect of frequency on semantic divergence. To this effect, we compute partial correlations, using polysemy as a covariate variable. Similarly, when computing correlations for polysemy, we set frequency as a covariate.

The results showing the correlations for each language pair are reported in Table 7.11. The values show a positive correlation, with values up to 0.33 (for Italian-French), suggesting that the frequency of a cognate word is related to the degree of semantic change it suffered, independently from polysemy.

We further try to understand the nature of the relationship between frequency and falseness. Previous studies (Hamilton et al. 2016) showed that prior frequency relates to subsequent semantic shift according to a power law. In our setup, we study the effect of previous semantic shift on the frequency of words. We model this relation by comparing the logarithm of the (average) frequency for a word pair with the falseness degree of the pair. To obtain the log-frequency, we use the Zipf frequencies provided by the Wordfreq library, which are computed as the base-10 logarithm of the number of times it appears per billion words. We first plot the log-frequency against the falseness degree, shown for Spanish-Portuguese in Figure 7.8.

Table 7.11: Correlations of frequency with falseness, controlling for polysemy.

	Es	Pt	It	Fr	Ro	En
Es		0.219	0.11	0.201	0.007	0.08
Pt	0.212		0.048	0.161	0.148	0.2
It	0.089	−0.007		0.334	0.129	0.083
Fr	0.188	0.117	0.323		0.194	0.3
Ro	0.062	0.148	0.147	0.271		0.229
En	0.161	0.242	0.083	0.315	0.163	

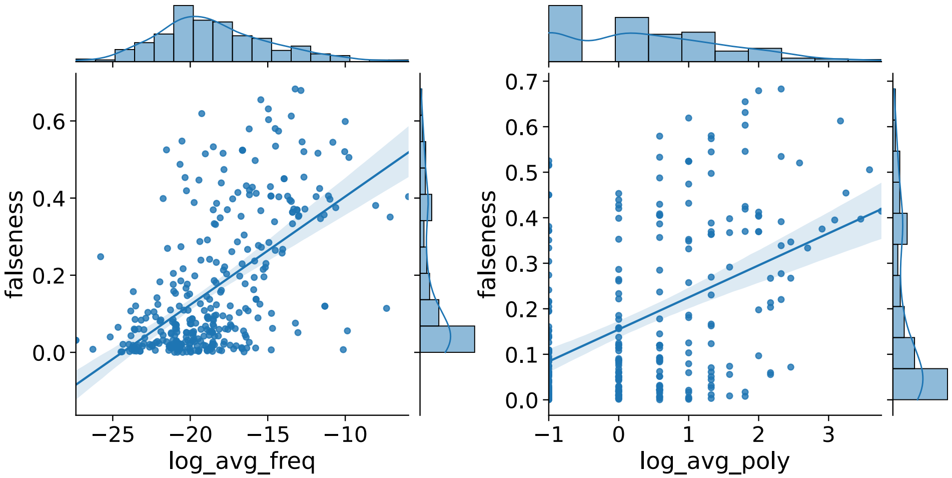


Figure 7.8: Falseness correlation with log-frequency and log-polysemy for Spanish-Portuguese.

We then try to find a function that describes in more precise terms the relationship between frequency and meaning shift, by fitting a polynomial curve of the following form:

$$\text{falseness} = a * (\log_{10}(\text{freq}))^b + c \tag{7.3}$$

where a, b and c are the parameters of the function.

Using the polynomial class of functions allows us flexibility in understanding the nature of the relationship (positive or negative, according to the sign of the main coefficient a), and its magnitude as measured by the size of the main coefficient and by the power coefficient b, while still being restrictive enough to facilitate computation using the limited training vocabulary in our dataset.

The complete list of the computed coefficients is shown in Table 7.12. We find that for most language pair there is a positive and superlinear relationship between the log-frequency score and the degree of falseness, with coefficients b generally close to 1, for all Romance language pairs, except for Romanian. For Romanian the power coefficients are higher across language pairs, associated with lower main coefficients, but their values less stable and less reliable, which we attribute to the lower coverage and quality of frequency scores, as well as to the inclusion of all cognates without filtering just the false friends. For pairs involving English, the algorithm generally fails to find a consistent relationship between the variables: it converges slower and is less stable, producing power coefficients that are very low, essentially resulting in an equation where the frequency variable is negligible.

It is interesting to compare our results with those of Hamilton et al. (2016), where the authors observe an inverse correlation between frequency and meaning shift: the more frequent words tend to change their meaning more slowly. Our experiments are set up to describe the phenomenon of semantic change from the opposite direction, measuring the expected frequency of a word that has undergone semantic change, and show the opposite effect: we find a positive relation – words that have diverged more in meaning tend to be more frequent.

4.2 Word polysemy and semantic divergence

For measuring *polysemy*, we make use of WordNet. In this way, the polysemy of a word can be defined as the number of synsets that it is part of in WordNet. As before, we have to exclude Romanian since it is not supported in WordNet (we assign a default polysemy score of 0 to all Romanian words). The polysemy score of a cognate pair is computed as the average between the polysemy scores of the two words involved.

Table 7.12: Optimal coefficients of polynomial describing function between frequency and falseness as falseness = $a * (\log_{10}(\text{freq}))^b + c$.

	Es			Pt		
	a	*b*	*c*	*a*	*b*	*c*
Es				0.04	1.27	−0.01
Pt	0.04	1.30	−0.03			
It	0.06	1.09	−0.07	0.08	0.80	−0.05
Fr	0.10	0.86	−0.12	0.07	0.91	−0.03
Ro	0.007	1.41	0.06	0.006	1.41	0.07
En	2.90	0.12	−3.10	201.8	0.001	−201.8
	It			Fr		
	a	*b*	*c*	*a*	*b*	*c*
Es	0.06	1.12	−0.08	0.25	0.58	−0.31
Pt	0.04	1.07	−0.01	0.07	0.96	−0.03
It				0.05	1.17	−0.004
Fr	0.10	0.82	−0.09			
Ro	0.01	1.06	0.05	0.004	1.86	0.07
En	493.1	0.05	−493.1	675.5	0.0005	−675.6
	Ro			En		
	a	*b*	*c*	*a*	*b*	*c*
Es	0.0003	3.11	0.08	580.3	0.0005	−580.4
Pt	0.0002	2.85	0.08	528.9	0.0005	−528.9
It	0.0003	3.17	0.08	463.8	0.0006	−463.8
Fr	0.0001	3.58	0.09	579.6	0.0005	−579.7
Ro				144.3	0.0007	−144.2
En	257.2	0.0003	257.1			

We perform similar experiments for polysemy, correlating the average degree of polysemy of the words in a cognate pair to the falseness of the pair, with frequency as a control variable. The obtained correlations, shown in Table 7.13, are noteworthy for most language pairs, with values as high as 0.47. Figure 7.8 shows the relationship between log-polysemy and falseness, which displays a clear linear trend.

Table 7.13: Correlations of polysemy with falseness, controlling for frequency.

	Es	Pt	It	Fr	Ro	En
Es		0.404	0.342	0.336	0.072	0.286
Pt	0.461		0.363	0.427	−0.004	0.305
It	0.305	0.383		0.412	0.019	0.087
Fr	0.341	0.413	0.479		−0.051	0.184
Ro	0.093	−0.01	−0.011	−0.049		−0.016
En	0.429	0.37	0.087	0.301	−0.062	

For Romance languages (with some isolated exceptions for pairs involving Romanian), polysemy proves to be strongly positively correlated with falseness, suggesting words which have undergone more semantic shift tend to be more polysemous. Previous studies (Hamilton et al. 2016) have found a positive relationship between polysemy and semantic change from the opposite perspective, showing that more polysemous words seem to suffer more semantic shift.

We further compute a polynomial that approximates the relationship between the log-polysemy score and falseness, following the general form:

$$\text{falseness} = a * \log_2(\text{polysemy})^b + c \tag{7.4}$$

where a, b and c are the coefficients to be found. In this case, all polysemy scores are non-negative integers. For words not found in WordNet, we replace the log-polysemy score with zeroes.

In the case of polysemy, we find a sublinear relationship between log-polysemy and falseness. Coefficients are listed in Table 7.14. The main coefficients are always positive, with the exception of one language pair (Romanian-English), and the power b is usually in the interval $(0.5, 1)$, for all languages, with the exception of Italian-English and several language pairs involving Romanian, where they are higher. We expect the results for Romanian to be less reliable in this case as well, since we lack polysemy scores for Romanian entirely (polysemy scores for language pairs involving Romanian rely entirely on the other language in the pair). In general, scores are more stable than in the case of the frequency-falseness law, including for English and for Romanian.

The positive relationship between falseness and both frequency and polysemy suggest words that have undergone semantic change tend to become more frequent as well as more polysemous, proportional to the degree of semantic shift, maintaining a consistent pattern across language pairs, especially among core Romance languages (Italian, French, Spanish, Portuguese). Overall, the higher power coefficients for frequency in relation to falseness (as compared to the case of polysemy) suggest that more pronounced shifts in meaning are associated with increases in frequency, as compared to polysemy where the meaning divergence associated with a change in polysemy is relatively milder. The more uniform coefficients in the laws relating polysemy and falseness across languages, as well as the higher partial correlation scores for polysemy, suggest a more consistent pattern of association between semantic divergence and polysemy (as compared to frequency).

4.3 Semantic fields of true cognates

In this final subsection, we present a brief qualitative analysis of true cognates across the Romance languages – the words that have preserved their meaning from their Latin etymon to the present day meaning, comparatively across languages. We use clustering methods to extract clusters of true cognates that can be interpreted to represent the different semantic fields of the words that have preserved their meaning throughout time. We show these examples as an initial attempt to better understand how the semantic change of words varies across semantic fields, comparatively across languages which diverged from Latin (based on the simplifying assumption that the extracted true cognates are inherited words) in different moments in history and different geographical, political and cultural contexts.

In order to obtain semantic clusters for each language, we first select the true cognates from our list of cognate sets, for language pairs consisting of Latin and a Romance language, using our algorithm (by selecting word pairs with null falseness scores). For each Romance language, we collect the vector representations of the extracted words in the corresponding embedding space. We then apply k-means clustering on this set of points to obtain 10 clusters, based on the cosine distance between the vectors, approximating the semantic distance between the extracted true cognates. For each resulted cluster, we find the centroid point,

Table 7.14: Optimal coefficients of polynomial describing function between polysemy and falseness as falseness = $a * (\log_2(\text{poly}))^b + c$.

	Es			Pt			It		
	a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
Es				0.10	0.99	0.11	0.11	0.78	0.11
Pt	0.10	0.93	0.10				0.07	1.02	0.11
It	0.10	0.56	0.12	0.07	0.73	0.11			
Fr	0.07	0.81	0.12	0.05	0.93	0.14	0.07	0.79	0.13
Ro	0.03	0.62	0.07	0.02	0.82	0.07	0.01	1.91	0.09
En	0.13	0.66	0.19	0.10	0.43	0.23	0.0004	3.60	0.32

	Fr			Ro			En		
	a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
Es	0.08	0.72	0.14	0.02	0.86	0.07	0.07	0.75	0.26
Pt	0.06	0.83	0.15	0.02	0.55	0.07	0.06	0.69	0.27
It	0.06	0.76	0.16	0.01	1.26	0.08	0.0004	3.60	0.32
Fr				0.004	2.23	0.08	0.02	0.98	0.30
Ro	0.004	2.18	0.09				−0.007	0.27	0.19
En	0.06	0.78	0.25	0.0001	4.99	0.16			

which we approximate with its closest word in the embedding space of that language (whether the resulting word is in our cognate set or not). In the end, we are left with 10 semantic clusters for each Romance language, each represented by a centroid word.

We show the resulting clusters in Appendix A. The first element of each line represents one cluster, containing the centroid word (rendered in bold letters), along with each word's translation in English (where they are meaningful), in decreasing order of the number of words in the cluster. Since not all centroid words are meaningful or representative, we also include on the same line, for each cluster, 4 representative words chosen manually from the true cognates belonging to the cluster, along with their translations.

It is interesting to see that there are a few semantic fields that occur consistently in each language: terms related to morality and justice, to medicine, and to chemistry or abstract mathematics. Some clusters are specific to certain domains that do not occur in each language, such as foods (in Portuguese), mechanical terms (in Portuguese and Romanian), or religious terms (in Romanian).

We should note a limitation of the method used to obtain the semantic clusters, stemming from the representation of Latin words. All word embedding representations used in this study were based on pre-trained FastText embeddings trained on Wikipedia (Conneau et al. 2017), including Latin embeddings. The uniformity of the text genres and topics across languages is in general an advantage to obtaining comparable embedding spaces, but in the case of Latin, the use of Latin Wikipedia, which consists of translated texts from other languages, might bias representations away from the original usages of the words in Latin, and towards their usages in the modern languages the texts were translated from, and consequently affect the detection of true cognates in relation to Latin. This drawback could be mitigated by using a corpus of original Latin texts instead for building the embeddings.

5 Conclusions

In this chapter, we proposed a method for computing the semantic divergence of cognates across languages, and showed how it can be used for computing language similarity, for measuring cross-lingual semantic change synchronically, as well as for practical applications including false friend detection and correction.

We defined a cross-lingual word similarity measure based on word embeddings and extended the pairwise metric to compute the semantic divergence across languages. Our results showed that Spanish and Portuguese are the closest languages, while Romanian is most dissimilar from Latin, possibly because it developed far from the Romance kernel. Furthermore, clustering the Romance languages based on the introduced semantic divergence measure resulted in a hierarchy that is consistent with the generally accepted tree of languages. When further including English in our experiments, we noticed that, even though most Latin words that entered English are probably borrowings (as opposed to inherited words), its similarity to Latin is close to that of the modern Romance languages. Our results shed some light on a new aspect of language similarity, from the point of view of cross-lingual semantic change.

We further showed how the introduced measures can be used for a practical application, and proposed a method for detecting false friends from cognate word pairs, distinguishing between two categories: hard false friends and soft false friends. Additionally, we built and made freely available a database of false friends in six languages, and evaluated it against WordNet and against a manually curated dataset of false friends, obtaining state of the art results. To the best of our knowledge, the published database is the largest public resource of its kind, both in terms of number of covered word pairs and considered languages. Additionally, the proposed method can be used to generate or detect pairs of false friends for any pair of languages, without requiring expensive manual work or dictionaries, but only large monolingual corpora to train word embeddings on, and small bilingual dictionaries to perform embedding space alignments.

We also proposed an algorithm for automatically correcting false friends, which to our knowledge is the first attempt in this direction. Along with false friends pairs, we published a falseness score for each pair, which can be used to customize the sensitivity to difference in meaning that defines a pair of false friends according to the application. We believe this resource can be very valuable for language learners, for example by incorporating false friends pairs in a tool to aid with language acquisition or text comprehension for non-natives, as well as for machine translation or other applications using natural language processing in a multilingual setting.

The unsupervised nature of the proposed algorithm also has the advantage of a high coverage of the vocabulary, unlike dictionary-based methods, which are prone to becoming outdated as language evolves. One disadvantage of our embedding-based algorithm is the lack of distinction between different senses of the same word. In the future it would be interesting to continue the study in the direction of considering also context-specific senses of words, in order to be able to better handle partial false friends, which are pairs of cognates which share meaning in some contexts and not in others. In the case of corrections as well, the method using word embeddings could be extended to provide false friend correction suggestions in a certain context (possibly by using the word embedding model to predict the appropriate word in a given context).

In the fourth section, we showed a new perspective for studying semantic change: comparing meaning of cognate words across languages. We showed how frequency and polysemy relate to semantic shifts of cognates across languages, demonstrating that both the frequency and polysemy of cognates positively correlate with their cross-lingual semantic shift, suggesting that semantic change, in the case of cognates, drives words to be both more frequent and more polysemous. Moreover, we found concrete functions that best approximate these relationships according to a power law, thus taking the first steps towards formulating statistical laws of cross-lingual semantic change.

In the future, including the proto-word in the analysis relating semantic shift to word properties (in this case, the Latin etymon) may give further insight into how cognates change their meaning, as well as allow the exploration of the reverse effect, that of the influence of word properties (frequency, polysemy, etc.)

on the magnitude of the subsequent semantic shift. Exploring alternative methods for obtaining multilingual word representations (through choosing a different training corpus or reducing the noise induced by embedding space alignment) would help to further strengthen our conclusions on multilingual language phenomena related to semantic change. Additionally, it would be interesting to further explain these correlations, as well as study other hypothesized laws of semantic change in a multilingual setting (such as the law of differentiation or parallel change, or the law of prototypicality), extending the study to other language families beyond the Romance cluster.

Appendix A Clusters of true cognates per language: centroids, representative words, and English translations.

- Spanish
  - **amoralidad (amorality)**, abstinente (abstinent), bueno (good), compasion (compassion), profeta (prophet)
  - –, ébano (ebony), equino (equine), playa (beach), estatuaria (statuary)
  - **reversibilidad (reversibility)**, abstracción (abstraction), ecuación (equation), exceso (excess), inductivo (inductive)
  - **contrar (counter)**, cazar (hunt), devastar (devastate), encender (ignite), salir (leave)
  - **cuotificación (quota)**, colectivo (collective), declaración (declaration), institución (institution), juez (judge)
  - **hiposalivación (hyposalivation)**, artrítico (arthritic), irritación (irritation), reumatismo (rheumatism), ulceración (ulceration)
  - **inmoralidad (immorality)**, abominable (abominable), disidente (dissident), indecente (indecency), perversidad (perversion)
  - **embrazadura (embracing)**, cicatriz (scar), hueso (bone), rotura (break), vibrar (vibrate)
  - **ahúma (smoke)**, bálsamo (balsam), freir (fry), arroz (rice), huevo (egg)
  - **higroscópico (hygroscopic)**, cáustico (caustic), evaporar (evaporate), fósforo (phosphorus), ópalo (opal)

- Portuguese
  - –, ébano (ebony), cobra (snake), hipódromo (hippodrome), vaqueiro (cowboy)
  - **pessoalização (personalization)**, acessível (accessible), conciliação (conciliation), educação (education), monarquia (monarchy)
  - **imoralidade (immorality)**, abjeto (abject), dissidente (dissident), incriminar (incriminate), promíscuo (promiscuous)
  - **amoralidade (amorality)**, admiração (admiration), autêntico (authentic), generosidade (generosity), monogamia (monogamy)
  - **pessoalizar (personalize)**, causar (cause), decidir (decide), justificar (justify), sugerir (suggest)
  - **pneumogástrico (pneumogastric)**, atrofia (atrophy), inflamação (inflammation), irritação (irritation), tosse (cough)
  - **irrotacional (irrotational)**, dispositivo (device), oscilação (oscillation), esférico (spherical), vibrar (vibrate)
  - **sêmola (semolina)**, agricultura (agriculture), bovino (bovine), forno (oven), suco (juice)
  - –, abstração (abstraction), convergir (converge), infinito (infinite), quadrilátero (quadrilateral)
  - **anidrita (anhydrite)**, alumínio (aluminium), emoliente (emollient), insolúvel (insoluble), viscoso (viscous)
- Italian
  - **moralizzare (moralize)**, apprendere (learn), cooperare (cooperate), evacuare (evacuate), presentare (present)
  - –, albore (dawn), complesso (complex), lupo (wolf), tempo (time)
  - **giovevole (beneficial)**, austero (austere), circospetto (circumspect), delicato (delicate), delinquente (delinquent)
  - **commiserazione (commiseration)**, affettazione (affectation), catastrofe (catastrophe), inspirazione (inspiration), emozione (emotion)
  - –, accelerazione (acceleration), ciclico (cyclic), deduttivo (deductive), eccesso (excess)
  - **espromissione (indivisibile)**, anticipazione (anticipation), comunicazione (communication), creazione (creation), emanazione (emanation)

  - **ulcerazione (ulceration)**, abortivo (abortion), cicatrice (scar), glaucoma (glaucoma), irritazione (irritation)
  - **insaporimento (flavor)**, coagulare (coagulate), fermentare (fermen), masticare (chewing), vegetare (vegetate)
  - **carbonatazione (carbonation)**, asfalto (asphalt), cristallino (crystalline), corrodere (corrode), ossidiana (obsidian)
  - **stilofaringeo (stylopharyngeal)**, vescica (bladder), coronale (coronal), polmone (lang), ventrale (ventral)

- French
  - **amoralité (amorality)**, arrogance (arrogance), conjugal (conjugal), émotion (emotion), inflexible (inflexible)
  - –, ambassade (embassy), convention (convention), éducation (education), gymnastique (gymnastics)
  - **apppliquer (apply)**, menacer (threat), combiner (combine), extraire (extract), jouer (play)
  - –, apostrophe (apostrophe), déductif (deductive), indication (indication), oeil (eye)
  - **bénédictionnaire (blessed)**, autographe (autograph), décalogue (decalogue), martyr (martyr), nobiliaire (nobiliary)
  - –, cyclique (cyclic), convexité (convexity), intersection (intersection), saturation (saturation)
  - **chapelure (breadcrumbs)**, chaud (warm), crème (cream), macération (maceration), spatule (spatula)
  - **entérotoxémie (enterotoxemia)**, clinique (clinical), convalescent (convalesent), immunité (immunity), médication (medication)
  - **testiculaire (testis)**, atrophie (atrophy), genou (knee), estomac (stomach), vertébré (vertebra)
  - **silicique (silicic)**, gélatine (gelatine), caustique (caustic), corroder (corrode), pigment (pigment)

- Romanian
  - **grăunțoasă (grainy)**, aramă (copper), cuc (cuckoo), plajă (beach), suc (juice)

- **senilitate (senility)**, anticipație (anticipation), enormitate (enormity), fascinație (fascination), intrigă (intrigue)
- –, mecanic (mechanical), accesibil (accessible), arc (arc), fluctuație (fluctuation)
- **incognoscibil (unknowable)**, adorabil (adorable), afective (affective), inaccesibil (inaccessible), invizibil (invisible)
- **limfadenopatie (lymphadenopathy)**, apoplecie (apoplecia), cicatrice (scar), letal (lethal), coagula (clot)
- –, beneficiar (beneficiary), comitet (committee), învăța (learn), proprietar (owner)
- **condamnabil (condemnable)**, calomnia (slander), discrimina (discriminate), nega (deny), infidel (unfaithful)
- **triclorurii (trichloride)**, aprinde (ignite), compozit (composite), cristalin (crystalline), cuaternar (quaternary)
- **pantocrator (pantocrator)**, altar (altar), cruce (cross), exorcist (exorcist), sacrilegiu (sacrilege)
- **perifraze (periphrases)**, apostrof (apostrophe), impersonal (impersonal), intranzitiv (intransitive), predicativ (predicative)

Acknowledgements

We thank the reviewers for their helpful and constructive comments. This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS - UEFISCDI, CoToHiLi project, number PN-III-P4-ID-PCE-2020-1544, within PNCDI III.

Abbreviations

BLI	Bilingual Lexicon Induction	Pt	Portuguese
Es	Spanish	Ro	Romanian
FF	false friends	Ro	Romanian
freq	frequency	UPGMA	Unweighted pair group method with arithmetic mean
Fr	French		
It	Italian		
La	Latin	WN	WordNet
poly	polysemy		

References

Aminian, Maryam, Mahmoud Ghoneim & Mona Diab. 2015. Unsupervised false friend disambiguation using contextual word clusters and parallel word alignments. In *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 39–48.

Artetxe, Mikel, Gorka Labaka & Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, 2289–2294.

Beinborn, Lisa & Rochelle Choenni. 2019. Semantic drift in multilingual representations. *arXiv preprint arXiv:1904.10820.*

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the ACL* 5. 135–146.

Bond, Francis & Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, volume 1: Long papers*, 1352–1362.

Campbell, Lyle. 1998. *Historical linguistics: An introduction.* Cambridge, MA: MIT Press.

Castro, Santiago, Jairo Bonanata & Aiala Rosá. 2018. A high coverage method for automatic false friends detection for Spanish and Portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial 2018*, 29–36.

Chen, Yanqing & Steven Skiena. 2016. False-friend detection and entity matching via unsupervised transliteration. *arXiv preprint arXiv:1611.06722.*

Ciobanu, Alina Maria & Liviu P. Dinu. 2014. Building a dataset of multilingual cognates for the Romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, 1038–1043.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer & Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087.*

Dinu, Liviu P. & Alina Maria Ciobanu. 2014. On the Romance languages mutual intelligibility. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, 3313–3318.

Dominguez, Pedro J. Chamizo & Brigitte Nerlich. 2002. False friends: Their origin and semantics in some selected languages. *Journal of Pragmatics* 34(12). 1833–1849.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer & Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of NetWordS 2015* (CEUR Workshop Proceedings 1347), 66–70. Pisa.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Eger, Steffen, Armin Hoenen & Alexander Mehler. 2016. Language classification from bilingual word embedding graphs. In *Proceedings of COLING 2016: Technical papers*, 3507–3518.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Heyman, Geert, Ivan Vulić & Marie Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proceedings of EACL 2017: Volume 1, long papers*, 1085–1095. ACL.

Inkpen, Diana, Oana Frunza & Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in French and English. In *Proceedings of RANLP 2005*, vol. 9, 251–257.

Koplenig, Alexander. 2016. *Analyzing lexical change in diachronic corpora*. Universität Mannheim. (Doctoral dissertation). http://nbn-resolving.de/urn:nbn:de:bsz:mh39-48905.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*, 1384–1397. Santa Fe: ACL.

Mikolov, Tomas, Quoc V. Le & Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Miller, George A. 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT press.

Mitkov, Ruslan, Viktor Pekar, Dimitar Blagoev & Andrea Mulloni. 2007. Methods for extracting and classifying pairs of cognates and false friends. *Machine Translation* 21(1). 29–53.

Nakov, Svetlin, Preslav Nakov & Elena Paskaleva. 2009. Unsupervised extraction of false friends from parallel bi-texts using the web as a corpus. In *Proceedings of RANLP 2009*, 292–298.

Patra, Barun, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley & Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 184–193.

Smith, Samuel L., David H. P. Turban, Steven Hamblin & Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Søgaard, Anders, Yoav Goldberg & Omer Levy. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of EACL 2017*, 765–774.

Søgaard, Anders, Sebastian Ruder & Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of ACL 2018: System demonstrations*, 778–788. ACL.

Sokal, Robert R. & Charles Duncan Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* 38. 1409–1438.

Speer, Robyn, Joshua Chin, Andrew Lin, Sara Jewett & Lance Nathan. 2018. *LuminosoInsight/wordfreq: V2.2*. DOI: 10.5281/zenodo.1443582.

St Arnaud, Adam, David Beck & Grzegorz Kondrak. 2017. Identifying cognate sets across dictionaries of related languages. In *Proceedings of EMNLP 2017*, 2519–2528.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278*.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 1–91. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040302.

Torres, Lianet Sepúlveda & Sandra Maria Aluísio. 2011. Using machine learning methods to avoid the pitfall of cognates and false friends in Spanish-Portuguese word pairs. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 67–76.

Uban, Ana, Alina Ciobanu & Liviu Dinu. 2019a. A computational approach to measuring the semantic divergence of cognates. In *Proceedings of CICLing 2019*.

Uban, Ana, Alina Maria Ciobanu & Liviu P. Dinu. 2019b. Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 161–166. Florence: ACL. DOI: 10.18653/v1/W19-4720.

Vulić, Ivan & Marie-Francine Moens. 2013. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of NAACL-HLT 2013*, 106–116. ACL.

Vulić, Ivan & Marie-Francine Moens. 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In *Proceedings of EMNLP 2014*, 349–362.

Vulić, Ivan & Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 2: short papers)*, 719–725. Beijing, China: Association for Computational Linguistics. DOI: 10.3115/v1/P15-2118.

Xu, Yang & Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual meeting of the Cognitive Science Society, CogSci 2015*. Pasadena.

Zhao, Shenjian & Zhihua Zhang. 2018. Attention-via-attention neural Machine translation. In *The Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018*.

Chapter 8

Structured representation of temporal document collections by diachronic linguistic periodization

Yijun Duan[a], Adam Jatowt[b] & Masatoshi Yoshikawa[c]
[a]National Institute of Advanced Industrial Science and Technology [b]University of Innsbruck [c]Kyoto University

Language is our main communication tool. Deep understanding of its evolution is imperative for many related research areas including history, humanities, social sciences, etc. as well as for effective temporal information retrieval. To this end, we are interested in the task of segmenting long-term document corpora into naturally coherent periods based on the embodied evolving word semantics. There are many benefits of such segmentation including better representation of content in long-term document collections and support for modeling and understanding semantic drift. We propose a two-step framework for learning time-aware word semantics and periodizing document archive. The effectiveness of our model is demonstrated on the New York Times corpus spanning from 1990 to 2016.

1 Introduction

Language is an evolving and dynamic construct. Awareness of the necessity and possibilities of large scale analysis of the temporal dynamics on linguistic phenomena has increased considerably in the last decade (Zhang et al. 2015, Yao et al. 2018, Tahmasebi et al. 2021). Temporal dynamics play an important role in many time-aware information retrieval (IR) tasks. For example, when retrieving documents based on their embeddings, one needs accurate representations of content by temporal embedding vectors.

Yijun Duan, Adam Jatowt & Masatoshi Yoshikawa. 2021. Structured representation of temporal document collections by diachronic linguistic periodization. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 261–285. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040316

It is intuitive that if an IR system is required to effectively return information from a target time period T_a, it may fail to do so if it is unable to capture the change in context between T_a and the current time, or just another time period in the past T_b. To which extent is the context of T_a different from that of T_b? Are there any turning points in the interval between T_a and T_b when a significant context change occurred, or do T_a and T_b belong to the same stage in the evolving process of language rather? The capability of answering such questions is crucial for effective IR systems when coping with time-aware tasks. However, to the best of our knowledge, the research problem of distinguishing key stages in the evolution's trajectory of language still remains a challenge in the field of temporal IR.

Traditionally, a language's diachrony is segmented into pre-determined periods (e.g., the "Old", "Middle" and "Modern" eras for English) (Schätzle & Booth 2019), which is problematic, since such an approach may yield results concealing the true trajectory of a phenomenon (e.g., false assumption on abrupt turning point about the data). Moreover, these traditional segments are very coarse and can be easily obscured and derived from arbitrary and non-linguistic features (Degaetano-Ortlieb & Teich 2018). Thanks to accumulated large amounts of digitized documents from the past, it is now possible to employ large scale data-driven analyses for uncovering patterns of language change. Thus, instead of blindly adopting a pre-determined periodization scheme, data-driven approaches, which reflect actual changes in the data, and which are able to achieve meaningful generalizations, can be applied. This can not only help with evolutionary linguistic studies by providing data-driven evidence, but could also support better understanding of variations in performance of diverse temporal IR systems on different periods of a temporal document collection. Furthermore, automatic periodization can be also beneficial for many less-researched languages for which there may not be a sufficient number of historical linguistics-oriented studies and findings.

In this study, we design a data-driven approach for segmenting a temporal document collection (e.g., a long-term news article archive) into natural, linguistically coherent periods, thanks to which we can both capture the features involved in diachronic linguistic change, as well as identify the time periods when the changes occurred. Our approach is generic and can be applied to any diachronic data set. The detected periods could then be applied in diverse temporal IR scenarios, such as temporal analog retrieval and archival document recommendation.

Our method is based on the computation of dynamic word embeddings needed to properly represent changing word semantics. Semantic senses of words are subject to broadening, narrowing, or other kinds of shifts throughout time. For instance, *Amazon* originally referred to mythical female warriors (in ancient Greek mythology), while it assumed a new sense of a large e-commerce company since the mid 1990s.

Additionally, different words may become conceptually equivalent or similar across time. For example, a music device *Walkman* played a similar role of mobile music playing device 30 years ago as *iPod* plays nowadays. The phenomenon of evolving word semantics is however rarely considered in the existing corpus periodization schemes.

In this paper, we structure document collections by periodizing the evolving word semantics embodied in the corpus. Specifically, for a long-term document corpus, our goal is to split the entire time span into several consecutive periods, where within the same period most words do not undergo significant fluctuations in term of their senses, while linguistic shifts are on the other hand relatively prevalent across different periods. In other words, a word is represented by a constant vector within one period, while it may have fairly different representations in different periods (see Figure 8.1).

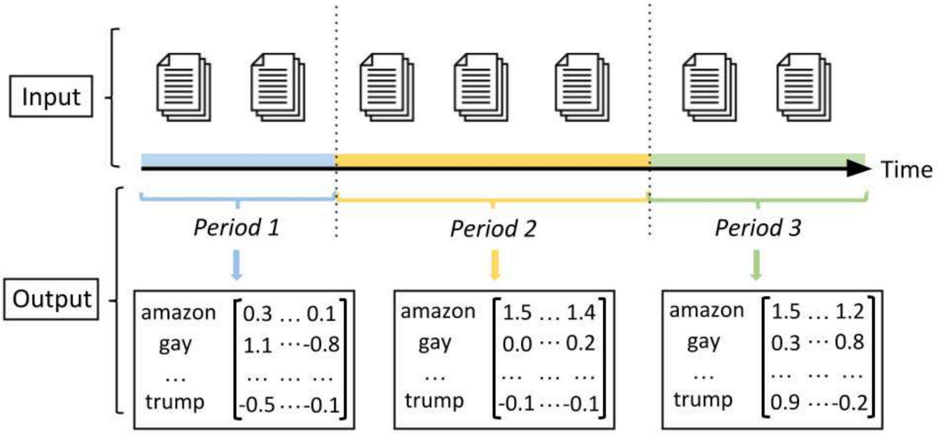


Figure 8.1: Conceptual view of our task. Our goal is to identify latent periods in the input document collection, such that word semantics are relatively stable within the same period (i.e., a word is represented by the same embedding vector), and major linguistic shifts exist between different periods (i.e., a word may be represented by fairly different vectors in different periods).

The problem of document collection periodization based on evolving word semantics is however not trivial. In order to solve this problem, we address the following two research questions:

a. How to compute temporal-aware word embeddings?

b. How to split the document collection based on learned word embeddings?

Our main technical contribution lies in a two-step framework for answering the above questions. First of all, we develop an anchor-based joint matrix factorization framework for computing time-aware word embeddings. More specifically, we concurrently factorize the time-stamped PPMI (positive pointwise mutual information) matrices, during which we utilize shared frequent terms (see Section 3) as anchors for aligning the word embeddings of each time frame to the same latent space. Furthermore, a block coordinate descent method is adopted to solve the learning model efficiently. Secondly, we formulate the periodization task as an optimization problem, where we aim to maximize the aggregation of differences between the word semantics of any two periods. To solve this problem, we employ three classes of optimization algorithms which are based on greedy splitting, dynamic programming and iterative refinement, respectively.

In the experiments, we use the crawled and publicly released New York Times dataset (Yao et al. 2018), which contains a total of 99,872 articles published between January 1990 and July 2016. We compare the performance of our models with existing competitive temporal word embedding methods, and corpus periodization methods, respectively. To demonstrate the quality of our learned temporal word embeddings, we focus on the task of searching for temporal analogs (see Section 5). To evaluate the periodization effectiveness, we construct the test sets by utilizing New York Times article tags (see Section 6), and evaluate the analyzed methods based on two standard metrics: Pk (Beeferman et al. 1999) and WinDiff (Pevzner & Hearst 2002) used in text segmentation tasks.

In summary, our contributions are as follows:

- From a conceptual standpoint, we introduce a novel research problem of periodizing diachronic document collections for discovering the embodied evolutionary word semantics. The discovered latent periods and corresponding temporal word embeddings can be utilized for many objectives, such as tracking and analyzing linguistic and topic shifts over time.

- From a methodological standpoint, we develop an anchor-based joint matrix factorization framework for computing time-aware word embeddings, and three classes of optimization techniques for document collection periodization.

- We perform extensive experiments on the New York Times corpus, which demonstrate the effectiveness of our approaches.

2 Problem definition

We start by presenting the formal problem definition.

2.1 Input

The input is a set of documents published across time. Each document is time-stamped and the whole text corpus spanning over a certain range of time is split into N basic time frames $(t_1, t_2, ..., t_N)$. The length of a time frame can be on the order of months, years, decades or centuries. Formally, let $D = \{D_1, D_2, ..., D_N\}$ denote the entire document set where $D_x, x = 1, ..., N$ represents the subset of documents belonging to the time frame t_x.

2.2 Task 1

Our first task is to find a d-dimensional embedding vector for each term in the overall corpus vocabulary $V = \{w_1, ..., w_{|V|}\}$,[1] for each time unit $t_i, i = 1, ..., N$, respectively. We denote by A_i the embedding matrix for t_i, whose j-th row represents the d-dimensional embedding vector of j-th term w_j in V. Thus A_i is of size $|V| \times d$.

2.3 Task 2

Based on Task 1, our second goal is to split the text corpus D into m contiguous, disjoint and coherent periods $\Theta = (P_1, P_2, ..., P_m)$ and compute their corresponding word embedding matrices $E_i, i = 1, ..., m$. Note that in this study the value of m is pre-defined. Each period $P_i = [\tau_b^i, \tau_e^i], i = 1, ..., m$ is expressed by two time points representing its beginning date τ_b^i and the ending date τ_e^i, with $\tau_b^1 = t_1$ and $\tau_e^m = t_N$. Let $L(\Theta) = (\tau_b^1, \tau_b^2, ..., \tau_b^m)$ denote the list of beginning dates of m periods, where $\tau_b^1 = t_1$. Notice that searching for Θ is equivalent to discovering $L(\Theta)$.

[1] The overall vocabulary V is the union of vocabularies of each time unit, and thus it is possible for some $w \in V$ to not appear at all in some time units. This includes emerging words and dying words that are typical in real-world news corpora.

3 Temporal word embeddings

In this section, we describe our approach for computing dynamic word embeddings (solving Task 1 in Section 2), that captures lexical semantic dynamics across time.

3.1 Learning static embeddings

The distributional hypothesis (Firth 1957) states that semantically similar words usually appear in similar contexts. Let v_i denote the vector representing word w_i, then v_i can be embodied in the co-occurrence statistics of w_i. In this study we first factorize the PPMI (positive pointwise mutual information) matrix for constructing static (i.e., time-agnostic) word embeddings, following previous works (Yao et al. 2018, Levy & Goldberg 2014, Hamilton et al. 2016).

For a corpus D with vocabulary V, the i, j-th entry of PPMI matrix (of size $|V| \times |V|$) is given by

$$\begin{aligned} \text{PPMI}_{i,j} &= \max\left\{\log_2\left(\frac{p(w_i, w_j)}{p(w_i)p(w_j)}\right), 0\right\} \\ &= \max\left\{\log_2\left(\frac{c(w_i, w_j) \cdot |D|}{c(w_i) \cdot c(w_j)}\right), 0\right\} \end{aligned} \tag{1}$$

where $p(w_i, w_j)$ represents the probability of words w_i and w_j co-occurring within a fixed-size sliding window of text, $c(w_i, w_j)$ counts the number of times that w_i and w_j co-occur, and $|D|$ is the total number of word tokens. Discarding the PPMI values under zero offers much better numerical stability (Yao et al. 2018).

For word vectors v_i and v_j, we should have $\text{PPMI}_{i,j} \approx v_i \cdot v_j$, thus such word vectors can be obtained through factorizing the PPMI matrix.

3.2 Learning dynamic embeddings

In order to compute the embedding matrices $E = E_1, ..., E_m$ for a given segmentation Θ on corpus D, we first construct the embedding matrix $A_i, i = 1, ..., N$ for each time unit. We denote PPMI_i the PPMI matrix for time frame t_i, thus temporal word embeddings A_i should satisfy $\text{PPMI}_i \approx A_i \cdot A_i^T$.

However, if A_i is constructed separately for each time unit, due to the invariant-to-rotation nature of matrix factorization these learned word embeddings A_i are non-unique (i.e., we have

$$\text{PPMI}_i \approx A_i \cdot A_i^T = (A_i W^T) \cdot (W A_i^T) = \tilde{A}_i \tilde{A_i^T}$$

for any orthogonal transformation W which satisfies $W^T \cdot W = I$). As a byproduct, embeddings across time frames may not be placed in the same latent space. Some previous works (Kulkarni et al. 2015, Hamilton et al. 2016, Zhang et al. 2015) solved this problem by imposing an alignment before any two adjacent matrices A_i and A_{i+1}, resulting in $A_i \approx A_{i+1}, i = 1, ..., N - 1$.

Instead of solving a separate alignment problem for circumventing the non-unique characteristic of matrix factorization, we propose to learn the temporal embeddings across time concurrently. Note that for a word, we desire its vector to be close among all temporal embedding matrices, if it did not change its meaning across time (or change its meaning to very small extent). Such words are regarded as "anchors" for connecting various embedding matrices, in our joint factorization framework.

Essentially, we assume that very frequent terms (e.g., *man*, *sky*, *one*, *water*) did not experience significant semantic shifts as their dominant meanings are commonly used in everyday life and by many people. This assumption is reasonable as it has been reported in many languages including English, Spanish, Russian and Greek (Lieberman et al. 2007, Pagel et al. 2007). We refer to these words as SFT, standing for SHARED FREQUENT TERMS. Specifically, we denote by A_i^{SFT} the $|V| \times d$ embedding matrix whose i-th row corresponds to the vector of word w_i in A_i, if w_i is a shared frequent term, and corresponds to zero vector otherwise, for a given time unit t_i. Our joint matrix factorization framework for discovering temporal word embeddings is then presented as follows (see Figure 8.2 for an illustration):

$$\begin{aligned} A_1, ..., A_N = \underset{\mathrm{A}}{\operatorname{argmin}} & \sum_{i=1}^{N} \left\| \mathrm{PPMI}_i - A_i \cdot A_i^T \right\|_F^2 \\ & + \alpha \cdot \sum_{i=1}^{N} \|A_i\|_F^2 + \beta \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left\| A_i^{SFT} - A_j^{SFT} \right\|_F^2 \end{aligned} \tag{2}$$

where the key smoothing term $\left\| A_i^{SFT} - A_j^{SFT} \right\|_F^2$ aligns shared frequent terms in all years, thus places word embeddings across time in the same latent space. The regularization term $\|A_i\|_F^2$ is adopted to guarantee the low-rank data fidelity for overcoming the problem of overfitting. Parameters α and β are used to control the weight of different terms to achieve the best factorization.

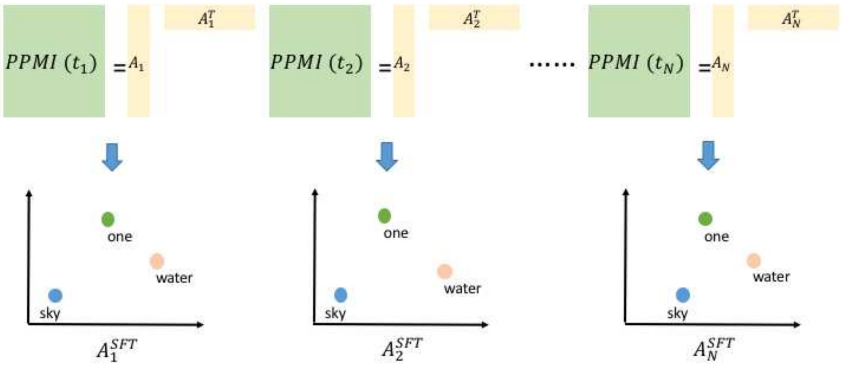

PPMI (t_1) = A_1 A_1^T
PPMI (t_2) = A_2 A_2^T
......
PPMI (t_N) = A_N A_N^T
one
water
sky
A_1^{SFT}
A_2^{SFT}
A_N^{SFT}

Figure 8.2: Illustration of our joint matrix factorization model. Shared frequent terms in all time frames $(t_1, t_2, ..., t_N)$ are aligned to similar positions, which places word embeddings across time in the same latent semantic space.

3.3 Optimization

The optimization problem in Equation (2) is not jointly convex to $A_i, i = 1, ..., N$, we first decompose the objective across periods, and solve for A_i by fixing other embedding matrices as constants at each step. The problem of optimizing A_i can be then formulated as follows:

$$\begin{aligned} A_i = \underset{\mathrm{A}}{\operatorname{argmin}}\ \Omega(A_i) = \underset{\mathrm{A}}{\operatorname{argmin}} \left\| \mathrm{PPMI}_i - A_i \cdot A_i^T \right\|_F^2 \\ + \alpha \cdot \|A_i\|_F^2 + \beta \cdot \sum_{j=1}^{N} \left\| A_i^{SFT} - A_j^{SFT} \right\|_F^2 \end{aligned} \tag{3}$$

Notice that $\Omega(A_i)$ is quartic in A_i, thus Equation (3) can not be optimized analytically. We then adopt the block coordinate descent (Tseng 2001) for iteratively minimizing $\Omega(A_i)$. Specifically, the gradient of $\Omega(A_i)$ with regard to A_i is given by

$$\frac{\partial \Omega(A_i)}{\partial A_i} = 2(A_i \cdot A_i^T - \mathrm{PPMI}_i + \alpha) \cdot A_i + 2\beta \cdot \sum_{j=1}^{N} (A_i^{SFT} - A_j^{SFT}) \tag{4}$$

where each above computation is of the order $O(\mathrm{nnz}(\mathrm{PPMI}_i)d + d^2 V)$ where $\mathrm{nnz}(\mathrm{PPMI}_i)$ is the number of non-zeros in the matrix.

4 Document collection periodization

In this section, we prescribe how to obtain the final periods (solving Task 2 in Section 2). We first introduce the scoring objective of linguistic periodization, then we study the effectiveness of three optimization approaches: (1) greedy algorithm based periodization, which searches for the best available boundary at each step; (2) dynamic programming based periodization, which is able to discover the optimal periods in a dynamic programming manner; (3) an iterative refinement scheme, which iteratively refines the boundaries for improving the performance of the greedy strategy.

4.1 Scoring

To frame the periodization problem as a form of optimization, having built a particular segmentation Θ, we now specify the way to quantify the quality of Θ, and then adopt different classes of techniques to optimize that scoring objective. In general, we prefer the embedding matrices of different periods to be characterized by high inter-dissimilarity. More explicitly, the objective $\mathrm{Obj}(\Theta)$ for an overall segmentation is given by aggregating the dissimilarity (expressed by the squared F-norm of the difference of two embedding matrices) between all pairs of period-specific embedding matrices, as follows:

$$\mathrm{Obj}(\Theta) = \mathrm{Obj}(L(\Theta)) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left\| E_i - E_j \right\|_F^2 \tag{5}$$

Here, m is the pre-defined number of periods. E_i is measured as the average of embeddings A_t for time unit t in period $P_i = [\tau_b^i, \tau_e^i]$, as follows:

$$E_i = \frac{1}{\tau_e^i - \tau_b^i + 1} \sum_{t=\tau_b^i}^{\tau_e^i} A_t \tag{6}$$

The segmentation that achieves the highest score of Equation (5) will be adopted.

4.2 Periodizing

4.2.1 Greedy algorithm based periodization

The greedy periodization algorithm is not guaranteed to reach the optimal splitting, however it offers significant computational benefit. At each step, it greedily inserts a new boundary (which is the beginning date of a new period) to the existing segmentation to locally maximize the objective function, until desired *m* periods are discovered. The process of greedy periodization is formulated as follows:

$$L(\Theta)^{i+1} = \underset{t_p \in [t_1, t_N], t_p \notin L(\Theta)^i}{\arg\max} \mathrm{Obj}(L(\Theta)^i \cup \left\{t_p\right\}) \tag{7}$$

where $L(\Theta)^i$ denotes the list of boundaries (or the beginning dates of periods) at the *i*-th step, and $L(\Theta)^0 = \{t_1\}$. The process of greedy algorithm based periodization is shown in Algorithm 1.

Algorithm 2: Greedy algorithm based periodization

```
input  : $L(\Theta)^0$; $m$
output : $L(\Theta)^{m-1}$
for $i \leftarrow 0$ to $m-2$ do
    $max\_score \leftarrow 0$;
    $next\_boundary \leftarrow 0$;
    for $t_p \leftarrow t_1$ to $t_N$ do
        ▷ Find the best local boundary;
        if $t_p \in L(\Theta)^i$ then
            continue
        end
        $score \leftarrow \mathrm{Obj}(L(\Theta)^i \cup \left\{t_p\right\})$;
        if $score > max\_score$ then
            $max\_score \leftarrow score$;
            $next\_boundary \leftarrow t_p$;
        end
    end
    $L(\Theta)^{i+1} \leftarrow L(\Theta)^i \cup \{next\_boundary\}$;
end
```

4.2.2 Dynamic programming based periodization

The core idea of dynamic programming based periodization is to break the overall problem into a series of simpler smaller segmentation tasks, and then recursively find the solutions to the sub-problems. By recursively solving the sub-problems optimally, the dynamic programming approach yields the globally optimal value of Equation (5). Let Θ_k^l denote the segmentation of the first l time slices of the entire time span into k periods. The computational process of dynamic programming based periodization is then expressed as follows:

$$L(\Theta_k^N) = \underset{l<N}{\arg\max}\ \mathrm{Obj}(L(\Theta_{k-1}^l) \cup t_{l+1}) \tag{8}$$

where $\Theta_1^l = [t_1, t_l]$ and $L(\Theta_1^l) = \{t_1\}, l = 1, ..., N$. In practice, though each of those sub-problems can be solved in one pass by storing their solutions in a memory-based data structure (array, map, etc), the dynamic programming approach can be costly to compute, compared to the greedy splitting, as shown below in Section 4.3. The process of dynamic programming based periodization is shown in Algorithm 2.

Algorithm 3: Dynamic programming based periodization

```
input   :L($\Theta_1^l$), $l = 1, ..., N$; $m$
output  :$L(\Theta_m^N)$
for $row \leftarrow 2$ to $m$ do
    for $col \leftarrow row$ to $N$ do
        ▷ Recursively find the solutions to the sub-problems;
        $max\_score \leftarrow 0$;
        $next\_boundary \leftarrow 0$;
        $subtask \leftarrow 0$;
        for $j \leftarrow row - 1$ to $col - 1$ do
            $score \leftarrow \mathrm{Obj}(L(\Theta_{row-1}^j) \cup \{t_{j+1}\})$;
            if $score > max\_score$ then
                $max\_score \leftarrow score$;
                $next\_boundary \leftarrow t_{j+1}$;
                $subtask \leftarrow j$;
            end
        end
        $L(\Theta_{row}^{col}) \leftarrow L(\Theta_{row-1}^{subtask}) \cup \{next\_boundary\}$
    end
end
```

4.2.3 Iterative refinement based periodization

The iterative refinement framework starts with the greedy segmentation. At each step after the best available boundary is found, a relaxation scheme which tries to adjust each segment boundary optimally while keeping the edges (i.e. adjacent boundaries) to either side of it fixed, is applied. This method can improve the performance of the greedy scheme, while at the same time retain its computational benefit to some extent. Let $L(\Theta)^i_G[j]$ denote the j-th element in $L(\Theta)^i$ after the i-th greedy search step, the iterative refinement process for finding $L(\Theta)^i[j]$ is shown as follows:

$$L(\Theta)^i[j] = \mathop{\arg\max}_{t_p \in \left(L(\Theta)^i[j-1], L(\Theta)^i[j+1]\right)} \mathrm{Obj}((L(\Theta)^i \setminus L(\Theta)^i_G[j]) \cup \left\{t_p\right\}) \qquad (9)$$

The process of this method is shown in Algorithm 3 below.

Algorithm 4: Iterative refinement based periodization

```
input  : $L(\Theta)^0$; $m$
output : $L(\Theta)^{m-1}$
for $i \leftarrow 0$ to $m-2$ do
    $next\_boundary, max\_score \leftarrow Greedy(L(\Theta)^i)$;
    $L(\Theta)^{i+1} \leftarrow L(\Theta)^i \cup \{next\_boundary\}$;
    for $j \leftarrow 1$ to $i$ do
        ▷ Iteratively refine the previous boundaries;
        $new\_boundary \leftarrow L(\Theta)^{i+1}[j]$;
        $t_{begin} \leftarrow L(\Theta)^{i+1}[j-1]$;
        $t_{end} \leftarrow L(\Theta)^{i+1}[j+1]$;
        for $t_p \leftarrow t_{begin}$ to $t_{end}$ do
            $score \leftarrow \mathrm{Obj}(L(\Theta)^{i+1} - L(\Theta)^{i+1}[j] \cup \left\{t_p\right\})$;
            if $score > max\_score$ then
                $max\_score \leftarrow score$;
                $next\_boundary \leftarrow t_p$;
            end
        end
        $L(\Theta)^{i+1} \leftarrow (L(\Theta)^{i+1} - L(\Theta)^{i+1}[j]) \cup \{new\_boundary\}$;
    end
end
```

4.3 Analysis of time complexity

For greedy periodization, it requires $m-1$ steps and the i-th step calls the scoring function Equation (5) $N-i$ times. In total, it is $Nm-N-m^2+m/2$. In the case of $N \gg m$, the greedy periodization algorithm takes $O(Nm)$. For dynamic programming based periodization, it requires $O(Nm)$ states and evaluating each state involves an $O(N)$ calling of Equation (5). Then the overall algorithm would take $O(N^2m)$. Finally, for iterative refinement based periodization, an upper bound on its time complexity is $O(\sum_{i=1}^{m-1}(N-i) * i) = O(Nm^2)$.

5 Embedding effectiveness

5.1 Datasets

News corpora, which maintain consistency in narrative style and grammar, form a good basis for studying language evolution. We perform the experiments on the New York Times Corpus, which has been frequently used to evaluate different researches that focus on temporal information processing or extraction in document archives (Campos et al. 2014). The dataset we use (Yao et al. 2018) is a collection of 99,872 articles published by the New York Times between January 1990 and July 2016. For the experiments, we first divide this corpus into 27 frames, setting the length of time unit to be 1 year. Stopwords and rare words (which have less than 200 occurrences in the entire corpus) were removed before experiments, following previous work (Zhang et al. 2015). The statistics of our dataset are shown in Table 8.1.

Table 8.1: Summary of the New York Times dataset

#Articles	#Vocabulary	#Word Co-occurences	#Time units	Range
99,872	20,936	11,068,100	27	Jan. 1990–Jul. 2016

5.2 Experimental settings

We describe next the parameters used in the experiments. For the construction of the PPMI matrix, the length of the sliding window and the embedding dimensions is set to be 5 and 50, respectively, following (Yao et al. 2018). During the training process, the values of parameters α and β (see Equation (2)) are set to be 20 and 100, respectively, as the result of a grid search. The selection of shared frequent terms (see Section 3.2) used as anchors is set to be the top 5% most popular words in the entire corpus excluding stopwords, as suggested by (Zhang et al. 2015).

5.3 Compared methods

We describe here the analyzed methods for learning temporal word embeddings.

Without transformation (Non-Tran): This method directly compares the vectors in different time without performing any transformation.

Linear transformation (LT) (Zhang et al. 2015): The embeddings are first trained separately for each year, and then are transformed by optimizing a linear transformation between adjacent years.

Orthogonal transformation (OT) (Hamilton et al. 2016): The embeddings are first trained separately for each year, and then are aligned by optimizing an orthogonal transformation between adjacent years.

Dynamic Word2Vec (DW2V) (Yao et al. 2018): The embeddings are trained based on PPMI matrices by minimizing the distance between embeddings in only adjacent years, without using SFTs.

The proposed model (this paper): The embeddings are jointly learned, by minimizing the difference between embeddings of shared frequent terms within the entire period.

We use the publicly available source code released by (Yao et al. 2018) for all baseline methods.[2]

5.4 Test sets

To demonstrate the effectiveness of our model, we focus on the task of searching for temporal analogs. We utilize 2 testsets (Yao et al. 2018) containing queries in the base time (e.g., *obama* in 2012) and their analogs in target time (e.g., *bush* in 2002). Testset 1 includes publicly recorded knowledge that for each year lists different names for a particular role (e.g., U.S. president),[3] and testset 2 consists

[2] https://github.com/yifan0sun/DynamicWord2Vec

[3] Note that we find several mistakes in this testset, such as (*pistons-1990, knicks-1999*) (the correct pair should be (*pistons-1990, spurs-1999*)). Then we manually correct them and use the corrected version for all analyzed methods in experiment.

of interesting concepts such as emerging technologies, brands and major events (e.g., *app* in 2012 can correspond to *software* in 1990: Yao et al. 2018). In total, there are 11,473 pairs of terms (query and its analog) used in our experiments.

5.5 Evaluation metrics

The MEAN RECIPROCAL RANK (MRR) is used for evaluating the search results for each learning model, which is computed as follows:

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \tag{10}$$

where $rank_i$ is the rank of a correct temporal analog at the i-th test, and N is the number of test pairs.

In addition, precisions @1, @5, @10 and @20 are also reported. Those metrics refer to the rates of tests in which the correct temporal analog was included in the top 1, 5, 10 and 20 results, respectively. All the values of used metrics fall into [0,1]. The higher the values are, the more effectively a model works.

5.6 Experimental results

Table 8.2: Performance of all analyzed models for learning dynamic word embeddings.

	Testset 1					Testset 2				
	MRR	P@1	P@5	P@10	P@20	MRR	P@1	P@5	P@10	P@20
Non-Tran	0.012	0.020	0.034	0.042	0.064	0.005	0.000	0.000	0.018	0.025
LT	0.137	0.118	0.232	0.267	0.355	0.038	0.021	0.065	0.146	0.219
OT	0.158	0.106	0.224	0.295	0.373	0.050	0.023	0.079	0.142	0.185
DW2V	0.422	0.331	0.549	0.619	0.703	0.144	0.076	0.220	0.382	0.487
Our model	0.454	0.348	0.563	0.651	0.740	0.157	0.082	0.255	0.406	0.520

Table 8.2 shows the scores for all the methods averaged on all the tested queries on testset 1 and testset 2, respectively. We first notice that the performance is extremely poor without transforming the contexts of queries. The correct answers in the Non-Tran approach are usually found at ranks > 1k which is in line with observations made by Zhang et al. (2015). On the other hand, both transformation-based methods LT and OT are helpful since they exhibit significantly better effectiveness compared to Non-Tran. This observation suggests little overlap in

the contexts of news articles which are separated by long time gaps, and that the task of temporal analog identification is quite difficult. Moreover, it is evident that learning the temporal embeddings across time by enforcing a global alignment is superior to following the "separately learning-and-aligning" pattern, since both DW2V and our approach outperform LT and OT significantly. Therefore, enforcing a global alignment is more effective for solving the temporal analog detection task.

Lastly, a closer look at Table 8.2 reveals that regardless of the type of evaluation metric, our model improves upon the performance of the state-of-the-art DW2V model. Specifically, our method improves DW2V model by 9.0% and 7.6% when measured using the main metric MRR on testset 1 and testset 2, respectively. The plausible reason is that DW2V does not differentiate words with stable meanings from words whose semantics are evolving, while such assumption may lead to a less precise learned representation of words. By injecting additional knowledge of shared frequent terms as anchors, our approach allows for only aligning embeddings of such stable words, and keeping the representation of other words exactly as their diachronic contexts express.

6 Periodization effectiveness

6.1 Datasets

We use the same news article datasets as described in Section 5.1.

6.2 Compared methods

We implemented below two types of periodization models as analyzed methods (proposed methods and baselines) in order to compare the periods they generate with the reference periods.

6.2.1 Baseline methods

We test four baselines as listed below.

Random: The segment boundaries are randomly inserted.

VNC (Gries & Hilpert 2012): A bottom-up hierarchical variability-based neighbor clustering (VNC) approach to periodization.

KLD (Degaetano-Ortlieb & Teich 2018): An entropy-driven approach which calculates the Kullback-Leibler Divergence (KLD) between term frequency features in text from temporally adjacent time periods to identify stages of language change.

CPD (Kulkarni et al. 2015): An approach which uses statistically sound change point detection (CPD) algorithms to detect significant linguistic shifts based on mean shift model.

6.2.2 Proposed methods

We list three proposed methods below (see Section 4.2).

These proposed methods adopt different strategies to optimize Equation (5), based on the temporal word embeddings obtained in Section 3.

G-WSE: Greedy periodization based on word semantic evolution.

DP-WSE: Dynamic programming periodization based on word semantic evolution.

IR-WSE: Iterative refinement based on word semantic evolution.

6.3 Test sets

As far as we know there are no standard testsets for New York Time Corpus. We therefore had to create test sets. Note that the collected news articles dataset is associated with some metadata, including title, author, publication time, and topical section label (e.g., *Science*, *Sports*, *Technology*) which describes the general topic of news articles. Such section labels could be used to locate the boundaries of word meanings.

Intuitively, if a word w is strongly related to a particular section s in year t, we associate w, s and t together and construct a $\langle w, s, t\rangle$ triplet. A boundary of w is registered if it is assigned to different sections in two adjacent years (i.e., both triplet $\langle w, s, t\rangle$ and $\langle w, s', t+1\rangle$ hold and $s \neq s'$).

More specifically, for each word w in the corpus vocabulary V we compute its frequency in all sections for each year t, and w is assigned to the section in which w is most frequent. Note that this word frequency information is not used in our learning model. In this study we utilize the 11 most popular and discriminative sections of the New York Times,[4] following previous work (Yao et al. 2018).

[4] These sections are *Arts*, *Business*, *Fashion & Style*, *Health*, *Home & Garden*, *Real Estate*, *Science*, *Sports*, *Technology*, *U.S.*, *World.*

Recall that parameter m denotes the number of predefined latent periods. For each different m, we first identify the set of words S_m characterized by the same number of periods. Then for each method and each value of m, we test the performance of such method by comparing the generated periods with the reference segments of each word in S_m, and then take the average. In this study, we experiment with the variation in the value of m, ranging from 2 to 10.

6.4 Evaluation metrics

We evaluate the performance of the analyzed methods with respect to two standard metrics: Pk (Beeferman et al. 1999) and WinDiff (Pevzner & Hearst 2002) used in text segmentation tasks. Both metrics use a sliding window of fixed size k over the document and compare the newly generated segments with the reference ones. Here k is generally set as follows (Beeferman et al. 1999):

$$k = \left\lfloor \frac{\#\mathit{time\ units}}{2 \cdot \#\mathit{periods}} \right\rfloor - 1 \tag{11}$$

Specifically, the Pk metric counts the number of disagreements on the probe elements as follows:

$$\mathrm{Pk} = \frac{1}{N-k} \sum_{i=1}^{N-k} [P_{\mathrm{hyp}}(i, i+k) \neq P_{\mathrm{ref}}(i, i+k)] \tag{12}$$

where N indicates the number of elements (in our case, the number of time units) and $P(i, i+k)$ is equal to 1 or 0 according to whether or not both element i and $i+k$ are recognized as being in the same segment in hypothesized segmentation P_{hyp} and reference segmentation P_{ref}. Since Pk metric has the disadvantage that it penalizes false positives more severely than false negatives (Alemi & Ginsparg 2015), the WinDiff metric was introduced. It is defined as follows:

$$\mathrm{WinDiff} = \frac{1}{N-k} \sum_{i=1}^{N-k} [W_{\mathrm{hyp}}(i, i+k) \neq W_{\mathrm{ref}}(i, i+k)] \tag{13}$$

where $W_{\mathrm{hyp}}(i, i+k)$ and $W_{\mathrm{ref}}(i, i+k)$ each count the number of boundaries between the time units i and $i+k$ in generated and reference segments, respectively. An error is registered if they are different. Both Pk and WinDiff give values in the range $[0, 1]$. They are equal to 0 if and only if an algorithm assigns all boundaries correctly. The lower the scores are, the better the algorithm performs.

6.5 Evaluation results

Tables 8.3 and 8.4 summarize the Pk and WinDiff scores for each method, respectively. Based on the experimental data we make the following observations.

- The proposed methods exhibit the overall best performance regarding both Pk and WinDiff metrics. More specifically, they outperform the best baseline under 7 of 9 predefined numbers of periods in terms of Pk, and 6 of 9 in terms of WinDiff. This demonstrates the effectiveness of our proposed periodization frameworks.

- Regarding baseline methods, Random achieves the worst performance as expected. CPD and KLD show competitive performance under certain settings. CPD gets two wins in terms of Pk, and KLD obtains three wins in terms of WinDiff.

- DP-WSE is the best performer among all three proposed periodization algorithms. It contributes 6 best performance in terms of Pk, and 5 in terms of WinDiff. Moreover, when compared to G-WSE and IR-WSE, DP-WSE shows a 3.79% and 3.24% increase in terms of Pk, and a 7.77% and 6.46% increase in terms of WinDiff, respectively. This observation is in good agreement with the theoretical analysis, which states that dynamic programming based segmentation sacrifices computational efficiency for the optimal splitting.

- The operation of iterative refinement indeed improves the performance of greedy periodization. However, the improvement is marginal: many results generated by IR-WSE are similar or identical to those from G-WSE.

7 Related work

7.1 Text segmentation

The most similar task to the document collection periodization is text segmentation. The task of text segmentation is formulated as splitting a chunk of text into meaningful sections based on their topic continuity, and it has many useful applications in information retrieval, text summarization, etc. Early text segmentation approaches include TextTiling (Hearst 1997) and the C99 algorithm (Choi 2000), which are based on some heuristics on text coherence using a bag

Table 8.3: Performance comparison using Pk (Lower scores indicate better performance).

Acronym	Number of periods								
	2	3	4	5	6	7	8	9	10
Random	0.467	0.474	0.545	0.522	0.542	0.480	0.480	0.480	0.539
VNC	0.385	0.253	0.249	0.290	0.282	0.302	0.302	0.294	0.303
KLD	0.385	0.278	0.244	0.270	0.276	0.278	0.284	0.290	0.304
CPD	0.238	0.234	0.246	0.260	0.282	0.263	0.249	0.299	0.338
G-WSE	0.115	0.201	0.248	0.282	0.300	0.310	0.312	0.292	0.303
DP-WSE	0.115	0.230	0.236	0.251	0.271	0.290	0.291	0.286	0.296
IR-WSE	0.115	0.201	0.244	0.279	0.300	0.304	0.312	0.292	0.303

Table 8.4: Performance comparison using WinDiff (Lower scores indicate better performance).

Acronym	Number of periods								
	2	3	4	5	6	7	8	9	10
Random	0.467	0.474	0.545	0.478	0.542	0.480	0.480	0.480	0.500
VNC	0.417	0.346	0.396	0.416	0.426	0.434	0.439	0.435	0.388
KLD	0.417	0.343	0.383	0.384	0.428	0.437	0.434	0.430	0.384
CPD	0.414	0.386	0.387	0.394	0.430	0.430	0.430	0.432	0.385
G-WSE	0.383	0.430	0.435	0.449	0.456	0.449	0.447	0.432	0.387
DP-WSE	0.383	0.336	0.387	0.403	0.423	0.422	0.430	0.431	0.388
IR-WSE	0.383	0.405	0.428	0.449	0.456	0.449	0.447	0.421	0.387

of words representation. Furthermore many attempts adopt topic models to inform the segmentation task, including Riedl & Biemann (2012), Du et al. (2013). Alemi & Ginsparg (2015) is a segmentation algorithm based on time-agnostic semantic word embeddings. Most text segmentation methods are unsupervised. However, neural approaches have also been explored for domain-specific text segmentation tasks, such as Sehikh et al. (2017). Many text segmentation algorithms are greedy in nature, such as Choi (2000), Choi et al. (2001). Moving beyond the greedy approach, some works search for the optimal splitting for their own objective using dynamic programming (Utiyama & Isahara 2001, Fragkou et al. 2004).

Apart from computer scientists, social scientists also have proposed a variety of methods to break a corpus into coherent sections. Related frameworks include those of Ruef (1999), Gries & Hilpert (2008), Alsudais & Tchalian (2016). Some studies are investigating the temporal topics in various corpora including news (Allan et al. 2001), historical documents (Duan et al. 2017) or scientific archives (Blei & Lafferty 2006, Wang & McCallum 2006).

7.2 Temporal word embeddings

How to best represent words with low-dimensional dense vectors has attracted consistent interest for several decades. Early methods are relying on statistical models (Lund & Burgess 1996, Blei et al. 2003), while in recent years neural models such as word2vec (Mikolov et al. 2013) and GloVE (Pennington et al. 2014) have shown great success in many NLP applications. Moreover, it has been demonstrated that both word2vec and GloVE are equivalent to factorizing the PMI matrix (Levy & Goldberg 2014), which motivates our approach.

The above methods assume word representation is time-agnostic. Recently some works explored computing time-aware embeddings of words, for analyzing linguistic change and evolution (Yao et al. 2018, Zhang et al. 2015, Hamilton et al. 2016, Kulkarni et al. 2015, Azarbonyad et al. 2017, Gonen et al. 2020). In order to compare word vectors across time most works ensure the vectors are aligned to the same coordinate axes, by solving the least squares problem (Zhang et al. 2015, Kulkarni et al. 2015), imposing an orthogonal transformation (Hamilton et al. 2016) or jointly smoothing every pair of adjacent time slices (Yao et al. 2018). Different from the existing methods, in this study we inject additional knowledge by using shared frequent terms as anchors to simultaneously learn the temporal word embeddings and circumvent the alignment problem.

8 Conclusion

This work approaches a novel task – diachronic document collection periodization. The special character of our task allows capturing evolutionary word semantics. The discovered latent periods can be an effective indicator of linguistics shifts and evolution embodied in analyzed diachronic textual corpora. To address the introduced problem we propose a two-step framework which consists of a joint matrix factorization model for learning dynamic word embeddings, and a well-defined optimization formulation for corpus periodization. For solving the resulting optimization problem we develop a series of effective algorithms. We perform extensive experiments to evaluate generated periods on the New York Times corpus spanning from 1990 to 2016, and show that our proposed methods perform favorably against diverse competitive baselines.

In the future, we plan to incorporate causal analysis for detecting correlated word semantic changes. We will also consider utilizing word sentiments in corpora periodization scenarios.

Acknowledgements

This research has been partially supported by MEXT Kakenhi grants (#17H01828, #18K19841 and #19H04215).

Abbreviations

CPD	change point detection
DP-WSE	dynamic programming based on word semantic evolution
DW2V	dynamic word2vec
G-WSE	greedy periodization based on word semantic evolution
IR-WSE	iterative refinement based on word semantic evolution
KLD	Kullback-Leibler Divergence
LT	linear transformation
Non-Tran	without transformation
OT	orthogonal transformation
VNC	variability-based neighbor clustering

References

Alemi, Alexander A & Paul Ginsparg. 2015. Text segmentation based on semantic word embeddings.

Allan, James, Rahul Gupta & Vikas Khandelwal. 2001. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, 10–18. DOI: 10.1145/383952.383954.

Alsudais, Abdulkareem & Hovig Tchalian. 2016. Corpus periodization framework to periodize a temporally ordered text corpus. In *Twenty-second Americas conference on information systems*. Red Hook, NY: Curran Associates, Inc.

Azarbonyad, Hosein, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx & Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of CIKM 2017*, 1509–1518. Singapore: ACM. DOI: 10.1145/3132847.3132878.

Beeferman, Doug, Adam Berger & John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning* 34. 177–210. DOI: 10.1023/A:1007506220214.

Blei, David M. & John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning*, 113–120.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3. 993–1022.

Campos, Ricardo, Gaël Dias, Alípio M. Jorge & Adam Jatowt. 2014. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)* 47(2). 1–41.

Choi, Freddy Y. Y. 2000. Advances in domain independent linear text segmentation. *arXiv preprint 0003083*.

Choi, Freddy Y. Y., Peter Wiemer-Hastings & Johanna D. Moore. 2001. Latent semantic analysis for text segmentation. In *Proceedings of the 2001 conference on empirical methods in Natural Language Processing*.

Degaetano-Ortlieb, Stefania & Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the second joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, 22–33.

Du, Lan, Wray Buntine & Mark Johnson. 2013. Topic segmentation with a structured topic model. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 190–200. ACL.

Duan, Yijun, Adam Jatowt & Katsumi Tanaka. 2017. Discovering typical histories of entities by multi-timeline summarization. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 105–114. DOI: 10.1145/3078714.3078725.

Firth, John R. 1957. *Papers in linguistics 1934–1951*. London: Oxford University Press.

Fragkou, Pavlina, Vassilios Petridis & Ath Kehagias. 2004. A dynamic programming algorithm for linear text segmentation. *Journal of Intelligent Information Systems* 23(2). 179–197.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah & Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 538–555.

Gries, Stefan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora* 3(1). 59–81.

Gries, Stefan Th. & Martin Hilpert. 2012. Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics. In Terttu Nevalainen & Elizabeth Closs Traugott (eds.), *The Oxford handbook of the history of English*. Oxford: Oxford University Press.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Hearst, Marti A. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* 23(1). 33–64.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Levy, Omer & Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, 2177–2185.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449(7163). 713.

Lund, Kevin & Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers* 28(2). 203–208.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint 1301.3781*.

Pagel, Mark, Quentin D. Atkinson & Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163). 717.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*, 1532–1543. Doha: ACL. DOI: 10.3115/v1/D14-1162.

Pevzner, Lev & Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics* 28(1). 19–36.

Riedl, Martin & Chris Biemann. 2012. Text segmentation with topic models. *Journal for Language Technology and Computational Linguistics* 27. 47–69.

Ruef, Martin. 1999. Social ontology and the dynamics of organizational forms: Creating market actors in the healthcare field, 1966–1994. *Social Forces* 77(4). 1403–1432.

Schätzle, Christin & Hannah Booth. 2019. DiaHClust: An iterative hierarchical clustering approach for identifying stages in language change. In *Proceedings* *of the 1st international workshop on computational approaches to historical language change*, 126–135.

Sehikh, Imran, Dominique Fohr & Irina Illina. 2017. Topic segmentation in ASR transcripts using bidirectional RNNs for change detection. In *2017 IEEE automatic speech recognition and understanding workshop (ASRU)*, 512–518. IEEE.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 1–91. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040302.

Tseng, Paul. 2001. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications* 109(3). 475–494.

Utiyama, Masao & Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, 499–506.

Wang, Xuerui & Andrew McCallum. 2006. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, 424–433. ACM. DOI: 10.1145/1150402.1150450.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao & Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 673–681.

Zhang, Yating, Adam Jatowt, Sourav Bhowmick & Katsumi Tanaka. 2015. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proceedings of ACL/IJCNLP 2015 (Volume 1: Long papers)*, 645–655. Beijing: ACL. DOI: 10.3115/v1/P15-1063.

Chapter 9

Lexical semantic change for Ancient Greek and Latin

Valerio Perrone[a], Simon Hengchen[b], Marco Palma[c], Alessandro Vatri[d], Jim Q. Smith[c,e] & Barbara McGillivray[f,e]

[a]Amazon [b]University of Gothenburg [c]University of Warwick [d]University of Oxford [e]The Alan Turing Institute [f]University of Cambridge

Change and its precondition, variation, are inherent in languages. Over time, new words enter the lexicon, others become obsolete, and existing words acquire new senses. Associating a word with its correct meaning in its historical context is a central challenge in diachronic research. Historical corpora of classical languages, such as Ancient Greek and Latin, typically come with rich metadata, and existing models are limited by their inability to exploit contextual information beyond the document timestamp. While embedding-based methods feature among the current state of the art systems, they are lacking in their interpretative power. In contrast, Bayesian models provide explicit and interpretable representations of semantic change phenomena. In this chapter we build on GASC, a recent computational approach to semantic change based on a dynamic Bayesian mixture model. In this model, the evolution of word senses over time is based not only on distributional information of lexical nature, but also on text genres. We provide a systematic comparison of dynamic Bayesian mixture models for semantic change with state-of-the-art embedding-based models. On top of providing a full description of meaning change over time, we show that Bayesian mixture models are highly competitive approaches to detect binary semantic change in both Ancient Greek and Latin.

1 Introduction

The study of lexical semantics in a diachronic perspective is of primary importance in lexicography, historical linguistics and other humanities. Capturing the

Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q. Smith & Barbara McGillivray. 2021. Lexical semantic change for Ancient Greek and Latin. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 287–310. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040318

semantic spectrum and historical change of individual words as well as performing large-scale diachronic analyses of the lexicon can help us answer important questions about the development of our culture and heritage. Recent research in natural language processing (NLP) has led to the development of computational models of lexical semantic change (LSC) which have the potential to add new insights to diachronic semantics. Most computational research in this area, however, has focussed on extant languages, and only a few attempts have been made to tackle this topic for ancient languages.

To address this, Perrone et al. (2019) introduced GASC (genre-aware semantic change), a novel dynamic Bayesian mixture model for semantic change, where the evolution of word senses over time is based on distributional information and on additional features, specifically genre. GASC can decouple sense probabilities and genre prevalence, a critical task in the case of genre-unbalanced languages corpora, and can incorporate different categorical metadata, such as author, geography, or style. GASC was developed for Ancient Greek and represents the state-of-the-art in computational modelling of lexical semantic change for this language.

On the other hand, word-embedding models have become the most common methods adopted in lexical semantic change detection (Kutuzov et al. 2018) and an open question remains regarding which methods are most appropriate for ancient languages. In this chapter, we offer the first systematic evaluation of Bayesian dynamic mixture models and word embeddings models for semantic change in Latin and Ancient Greek. These ancient languages provide insightful test cases of automatic lexical semantic change for several reasons. First, as in many other languages, a large number of Latin and Ancient Greek words are polysemous (Clarke 2010), and polysemous words offer us a chance to study semantic variation, particularly across genres, and its relation to semantic change (Leiwo et al. 2012). Also, the literary traditions of these two languages have rich transcribed high-quality corpora covering a large number of literary genres. Moreover, they offer the opportunity to test the performance of different methods on use data spanning several centuries. Finally, we can rely on the scholarship of these languages to validate our computational systems. Our code is freely available at https://git.io/Jqe7U.

The word *mus* is an example of polysemous word (it can mean ‘mouse’, ‘mussel’ or ‘muscle’). The variation in the distribution of meanings over time per genre is displayed in Figure 9.1. In this graph, lines represent the percentage of the occurrences of the target word in a literary genre across centuries, while bars represent the percentage of the occurrences of a specific sense of *mus* across centuries. When the trend in any line agrees with the one for any set of bars (for

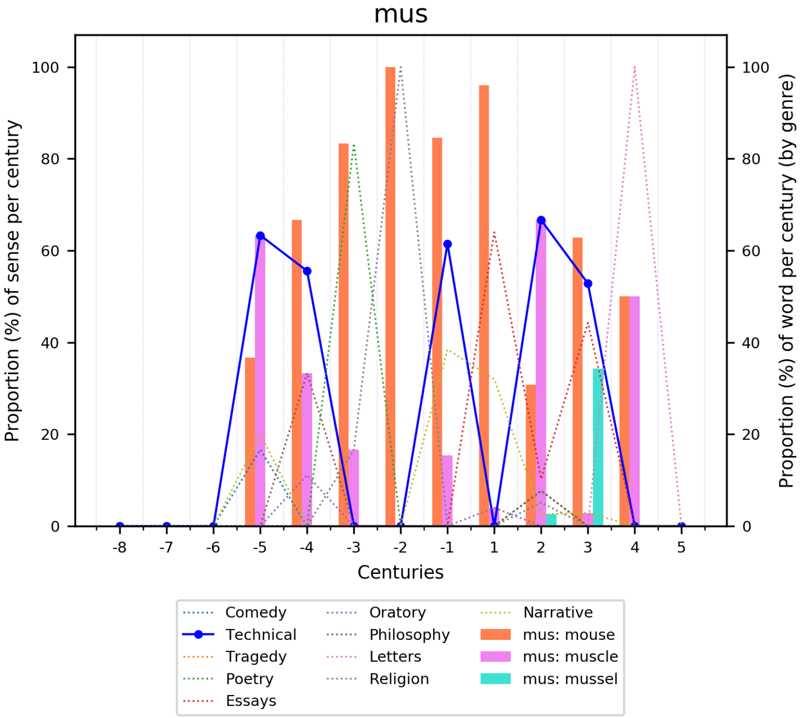


Figure 9.1: Distribution of *mus* 'mouse'/'muscle'/'mussel' by genre vs its senses over time (Perrone et al. 2019). Lines track *mus* proportions in each genre and century, while bars show the *mus* occurrence proportions with each sense and century.

instance, the distribution of 'muscle' over time tracks the blue line corresponding to the distribution of *mus* in technical genres), there might be evidence of genre-related changes.

In technical texts, we expect polysemous words to have a technical sense ('muscle' in the case of *mus*). On the other hand, in works more closely representing general language (comedy, oratory, historiography) we expect words to appear in their more concrete and less metaphorical senses ('mouse' or 'mussel' in the case of *mus*), although we cannot always assume that the same distribution holds in a number of other genres, such as philosophy and tragedy.

2 Related work

In recent years, NLP research has made great advances in the area of semantic change detection and modelling, with methods ranging from topic-based models (Boyd-Graber et al. 2007, Cook et al. 2014, Lau et al. 2014, Wijaya & Yeniterzi 2011, Frermann & Lapata 2016), to graph-based models (Mitra et al. 2014, 2015, Tahmasebi & Risse 2017), and word embeddings (Kim et al. 2014, Basile & McGillivray 2018, Kulkarni et al. 2015, Hamilton et al. 2016, Dubossarsky et al. 2017, Tahmasebi 2018, Rudolph & Blei 2018, Jatowt et al. 2018, Dubossarsky et al. 2019), to cite but a few.[1] However, models used in previous work are purely based on words' lexical distribution information and do not account for language variation features such as text type or genre. One reason for this is that genre-balanced corpora (such as COHA in Davies 2012) or single-genre corpora (such as newspapers, or Twitter, cf. e.g. Shoemark et al. 2019) are typically used. However, the strong role played by such factors in determining the sense of a word in context has been acknowledged in NLP research at least since Gale et al. (1992)'s idea of "one sense per discourse", according to which polysemous words tend to display the same sense in the same discourse. This principle has been widely adopted in word sense disambiguation research, with some more recent adaptations such as "one sense per Wikipedia category" (Scarlini et al. 2020).

Semantic change in ancient languages, especially on a large scale and over a long time period, is an under-explored research area. Previous work has mainly been qualitative in nature, due to the complexity of the phenomenon (cf. e.g. Leiwo et al. 2012, Clackson 2011). Some work has been done on training word embeddings on Ancient Greek (Rodda et al. 2019) and Latin (Sprugnoli et al. 2019) corpora, but not in a diachronic perspective. With the exception of a few works (Bamman & Crane 2011, Eger & Mehler 2016, Rodda et al. 2016, Perrone et al. 2019, McGillivray et al. 2019), two of which this chapter is based on and completes, no previous work has focussed on ancient languages.[2]

Recent work on languages other than English is rare but exists: Falk et al. (2014) use topic models to detect changes in French and Hengchen (2017) uses similar methods to tackle Dutch. Cavallin (2012) and Tahmasebi (2018) focus on Swedish, with the comparison of verb-object pairs and word embeddings, respectively. Zampieri et al. (2016) use SVMs to assign a time period to text snippets in Portuguese, and Tang et al. (2016) work on Chinese newspapers using S-shaped models. Most work in this area focusses on simply detecting the occurrence of semantic change, while Frermann & Lapata (2016)'s system, SCAN, takes into

[1] For an overview of the NLP literature, we refer to Tahmasebi et al. (2018), and Kutuzov et al. (2018) for a focus on neural embeddings. For an overview of the existing challenges in modelling and detecting semantic change, we refer to Hengchen et al. 2021 [this volume].

[2] To this list, we add the very recent SemEval 2020 Task 1 shared task on unsupervised lexical semantic change detection (Schlechtweg et al. 2020), which had Latin as one of its four target languages.

account synchronic polysemy and models how the different word senses evolve across time. More recently French has been further tackled by Jawahar & Seddah (2019), Frossard et al. (2020) and Montariol & Allauzen (2020), and German has been the focus of extensive work (Schlechtweg et al. 2017, 2018, 2019, 2020).

Our work bears important connections with the topic model literature. The idea of enriching topic models with document-specific author meta-data was explored in Rosen-Zvi et al. (2004) for the static case. Several time-dependent extensions of Bayesian topic models have been developed, with a number of parametric and nonparametric approaches (Blei & Lafferty 2006, Rao & Teh 2009, Ahmed & Xing 2012, Dubey et al. 2013, Perrone et al. 2017). In this chapter, we transfer such ideas to semantic change, where each datapoint is a bag of words associated to a single sense (rather than a mixture of topics). Excluding cases of intentional ambiguity, which we expect to be rare, we assume that there are generally no ambiguities in a context, and each word instance maps to a single sense. We acknowledge that this assumption can be seen as going against historical semantics literature (e.g. Traugott & Dasher 2001) which states that variation in context is the seed of semantic change.

3 The corpora

In order to conduct our experiments, we made use of two large diachronic corpora of Latin and Ancient Greek: LatinISE (McGillivray & Kilgarriff 2013) for Latin and the Diorisis Annotated Ancient Greek Corpus (Vatri & McGillivray 2018) for Ancient Greek. Our models require genre information. Genre-annotated corpora are not particularly common in NLP, where most tasks rely on specific genres (e.g. Twitter) or on genre-balanced corpora such as COHA (Davies 2002), but they are more prevalent within the humanities, and especially Classics. Additionally, research on automated genre identification has been flourishing for decades (e.g. Kessler et al. 1997), making the need for genre information in a potential corpus not as much of a hindrance as it could be thought.[3]

The Diorisis Annotated Ancient Greek Corpus contains 820 texts spanning between the beginnings of the Ancient Greek literary tradition (8th century BCE) and the 5th century CE. It is lemmatized and part-of-speech-tagged and contains 10,206,421 word tokens. Diorisis is the largest openly available annotated corpus

[3] While the influence of genre has been extensively studied in historical linguistics (see, for example, the extensive work by Biber & Finegan 1989), we use in this chapter a slightly different notion of *genre*: literary genre, as defined by classicists.

of Ancient Greek. The corpus covers a number of Ancient Greek literary and technical genres: poetry (narrative, choral, epigrams, didactic), drama (tragedy, comedy), oratory, philosophy, essays, narrative (historiography, biography, mythography, novels), geography, religious texts (hymns, Jewish and Christian Scriptures, theology, homilies), technical literature (medicine, mathematics, natural science, tactics, astronomy, horsemanship, hunting, politics, art history, rhetoric, literary criticism, grammar), and letters.

The LatinISE corpus (McGillivray & Kilgarriff 2013) covers 1,274 texts from between the beginnings of the Latin literary tradition (2nd century BCE) and the contemporary era (21st century CE). It has been automatically lemmatized and part-of-speech tagged. A domain expert manually added genre information for the following genres: comedy, essays, law, letters, narrative, oratory, philosophy, poetry, Christian, technical, tragedy. All Christian writings (including letters and poems) were assigned the genre Christian. This excludes philosophical but not theological or ecclesiological treatises composed by Christian writers.

4 Bayesian semantic change models

4.1 Domain knowledge elicitation

While NLP provides powerful tools to analyse texts, a central challenge is to ensure that outputs are explainable and that new discoveries can be placed within the context of the current state of the art in specific disciplines where NLP methods are applied. Bayesian methods have proved very useful within scientific modelling to incorporate domain explanations. In the Bayesian setting, expert judgements can be embedded directly into a probabilistic framework in the form of a prior. For instance, if historians know that a certain sense was popular in a given century, this information can be directly encoded into the model by changing the prior probability distribution for that sense. Data can then be analysed from these belief statements and a prior to posterior analysis performed, which helps domain experts adjust their beliefs in the light of the new available information (see for example Smith 2010, O'Hagan & Oakley 2014). These new outputs will be consistent with the explanations embedded within the probabilistic model, making results interpretable.

The challenge of applying Bayesian reasoning within the humanities is that typically domain experts have not been trained to reason probabilistically. Therefore, it is not possible to ask domain experts to provide direct probabilistic inputs to the Bayesian model. What it is possible instead is to elicit structural information, which can take a wide range of forms depending on the domain (Wilkerson

& Smith 2018). These structural models can usually be represented by a graph (i.e., as a set of nodes and connecting arcs) which captures the fundamental entities and their relationships. For example, an expert may know that a certain author predominantly uses *mus* to mean 'mouse'. The Bayesian modeller can then simply introduce a new node representing the author and condition the probability of using senses to the author variable. Once the graphical model is in place, we let the data quantify a joint probability model.

This work leverages the Bayesian network, one of the most developed structural models of this type. This structure embeds simple assertions about what measurements might be informative, in a way described by Korb & Nicholson (2009), Smith (2010), and Pearl (2009). Working backwards from the properties of the object of interest, we produce sequentially a collection of direct and indirect influences across the whole domain. The composite of the relationships can then be expressed by a single graph, called a plate diagram (see Figure 9.2 for such a plate diagram of our model). This plate diagram determines the factorisation of the corresponding probability density over these measurements.

We aimed to apply these structural elicitation techniques to the study of semantic variation and change in Latin and Ancient Greek. From discussions with Ancient Greek and Latin experts who have extensive experience with the corpora at hand, it emerged that one of the main drivers of this variation was the particular genre of the text. For instance, in works more closely related to general language (i.e. non-specialised, or non purely poetic language), such as comedy or historiography, we expect words to appear in their concrete and less metaphorical senses. The Ancient Greek word *mus* within a technical text would more likely mean 'muscle', while in narrative texts the meaning would more likely be 'mouse'. Such variations were believed to abound within the studied corpus. Since both the genre of texts was known to vary over time and text preservation to the current date depended on genre, any analysis which ignored genre might deduce a spurious change in overall meaning simply explainable from drifts in genre and selection effects influencing preservation. Having elicited this domain judgement, it was clear how to proceed. We simply modified the structure of our Bayesian model by adding genre as an additional observable variable (or node in the plate diagram). Conditioning on the observed genre, we could then have a specific distribution over senses accounting for genre-specific word usage patterns. Details of the model are given in the next section.

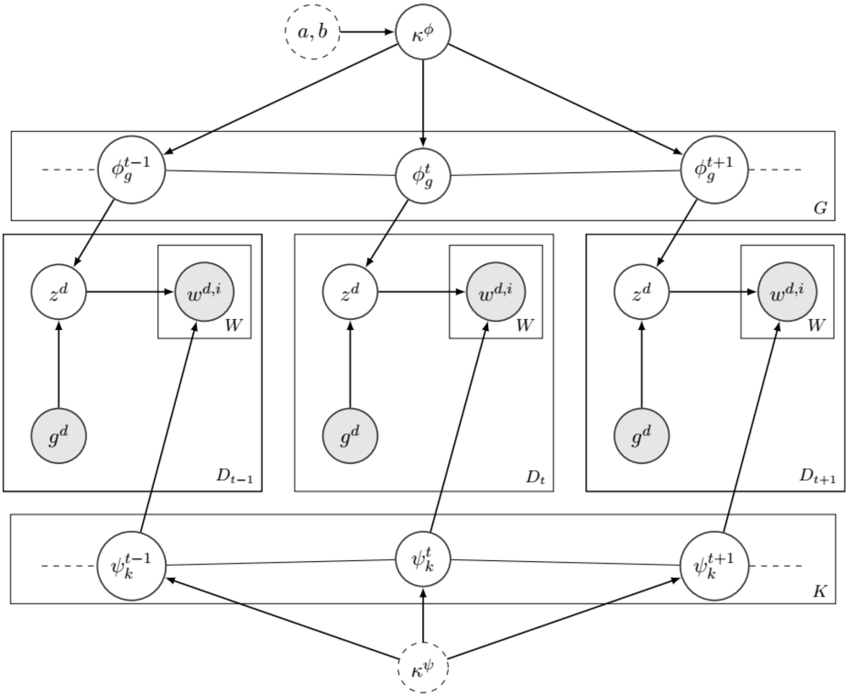


Figure 9.2: GASC plate diagram with three time periods (Perrone et al. 2019).

4.2 Genre-aware semantic change

A successful approach to model semantic change in Ancient Greek is GASC (Perrone et al. 2019). The starting point is a lemmatised corpus pre-processed into a set of text snippets of size W, each containing an instance of the word under study (referred to as “target word” in the remainder). The inferential task is to detect the sense associated to the target word in the given context, and describe the evolution of sense proportions over time.

We briefly summarise the generative model for GASC (illustrated by the plate diagram in Figure 9.2), which extends SCAN (Fremmann & Lapata 2016) to be genre-aware and is described in detail in Perrone et al. (2019). First, suppose that throughout the corpus the target word is used with K different senses, where we define a sense at time t as a distribution ψ_k^t over words from the dictionary. This statistical definition of sense is necessary to formalize the generative models presented in this work, and will be used throughout the rest of the paper.[4] These

[4] We follow the terminology adopted by Frermann (2017) and represent the meaning of a word

distributions are used to generate text snippets by drawing each of their words from the dictionary based on a multinomial distribution. Based on the intuition that each genre is more or less likely to feature a given sense, we assume that each of G possible text genres determines a different distribution over senses. Each observed document snippet is then associated with a genre-specific distribution over senses $\phi^t_{g^d}$ at time t, where g^d is the observed genre for document d. Conditioning on the observed genre yields a specific distribution over senses accounting for genre-specific word usage patterns. Word and sense distributions evolve over time with Gaussian changes, allowing for smooth transitions.

The model can be applied to different inferential goals: we can focus on the evolution of sense probabilities or on the changes within each sense. As we define a sense at time t as a probability distribution over words from the dictionary, this means that we can either choose to focus on the change of the sense probability over time or on the change in probability of the words characterising that sense. For each of these aims, we can use several hyperparameter combinations for K^{ϕ}, which is drawn from the prior distribution as determined by a and b, and K^{ψ}. To effectively detect semantic change points, the sense probabilities should not vary too smoothly over time and the bag of words should remain stable throughout the time periods.[5] For these reasons, we set the hyperparameters $a = b = 1, K^{\psi} = 100$ (equivalent to Setting 3 in Perrone et al. 2019). In particular, the hyperparameter K^{ψ} controls the homogeneity of the bag of words within the same sense and allows the emergence of new senses. This hyperparameter setting is used for SCAN and GASC on Latin, as well as for SCAN on Ancient Greek. For GASC on Ancient Greek, where the corpus size and the number of occurrences of target words is split between genres, the set of hyperparameters used is $a = 7, b = 3$, $K^{\psi} = 10$, as in Fremmann & Lapata (2016).

Further quantities to be set before running the Bayesian models are the number of iterations and the window size parameter W. The first runs of the Bayesian models usually show high variability in the results before convergence occurs; therefore, it is necessary to use a large number of iterations, especially for small sample sizes. For posterior inference we discard the first 100 iterations (burn-in period) and we run 2,500 iterations for models on Latin and 10,000 for models on

as a set of senses, each of which captures “an internally coherent aspect of its meaning, and is characterized through a set of words that are associated with that sense” (Frermann 2017: 173). We also assume that each instance of a target word in the corpus refers to one and only one sense.

[5] We acknowledge that the task of detecting change points is a drastically reduced view of semantic change. Nonetheless, as further explained in Section 6, this is required for ground truth evaluation.

Ancient Greek. The window size parameter W, namely the number of words to the left and right of an instance of the target, must also be carefully chosen not to introduce noisy irrelevant contextual words. Following Fremann & Lapata (2016) and Perrone et al. (2019), we fix the window size W to 5 for all methods and languages.

For posterior inference, we extend the blocked Gibbs sampler proposed in Frermann & Lapata (2016). The full conditional is available for the snippet-sense assignment, while to sample the sense and word distributions we adopt the auxiliary variable approach from Mimno et al. (2008). The sense precision parameters are drawn from their conjugate Gamma priors. For the distribution over genres we proceed as follows. First, sample the distribution over senses ϕ^t_g for each genre $g = 1, \dots, G$ following Mimno et al. (2008). Then, sample the sense assignment conditioned on the observed genre from its full conditional:

$$p(z^d \mid g^d, \mathbf{w}, t, \phi, \psi) \propto p(z^d \mid g^d, t) p(\mathbf{w} \mid t, z^d) = \phi^t_g \prod_{w \in \mathbf{w}} \psi^{t, z^d}_w .$$

This setting easily extends to sample genre assignments for tasks where, for example, some genre metadata are missing.

5 Embedding-based models

Neural-based word vectors are currently the most used representations in LSC. While skip-gram with negative sampling (SGNS, Mikolov et al. 2013) type embeddings have the limitation that they conflate senses of a word to a single vector representation, they currently perform better than other approaches, including contextual models such as ELMO (Peters et al. 2018) and BERT (Devlin et al. 2019), as reported by Schlechtweg et al. (2020).

In this chapter, we compare GASC and SCAN to the current state of the art in LSC (temporal referencing (TR), Dubossarsky et al. 2019), as well as with the oft-used combination of independently-trained SGNS models that are subsequently aligned using Orthogonal Procrustes (OP) proposed by Hamilton et al. (2016). Both models are very similar and rely on the same algorithm with the difference that TR, in which target words have different representations for every time bin but context words do not, has repeatedly been shown to produce much less noisy models (e.g. in Cassotti et al. 2020, Zamora-Reina & Bravo-Marquez 2020). In order to compare their performance with GASC, we train models on the whole corpus (“NAIVE”), as well as on genre subcorpora. For Ancient Greek we train models on Technical, NOT-technical, Narrative, NOT-narrative subcorpora, while Latin is divided between Christian and NOT-Christian.

6 Evaluation

Evaluating models tackling lexical semantic change is notoriously challenging. Schlechtweg et al. (2020) present the first shared task on unsupervised lexical semantic change detection, organized as part of the SemEval 2020 workshop. The task focusses on two subtasks: a binary classification task (for a set of target words, decide which words lost or gained senses between a time period t_1 and a time period t_2) and a ranking subtask (rank the same set of target words according to their degree of lexical semantic change between t_1 and t_2). The task provides gold standard sets for three extant languages (English, German, and Swedish) and one extinct language (Latin).

The Latin gold standard reflects the lexical semantic change in a portion of the Latin lexicon from Before the Common Era (BCE) and the Common Era (CE). For each of 40 lemmas selected from the corpus, expert annotators annotated 30 sentences extracted from a subcorpus of LatinISE consisting of texts from BCE, and 30 sentences from CE. For each sentence, the annotators selected one of four values (4 – identical, 3 – closely related, 2 – distantly related, 1 – unrelated) for each dictionary sense of the lemma, indicating the degree of similarity between the usage of the lemma reflected in the sentence and the dictionary sense. This choice of design implying that every target word has a closed set of possible senses corresponding to those listed in their respective dictionary entries is justified in the original paper.

The annotated data was analysed with a clustering technique that identified 26 lemmas as "changed" lemmas (meaning that they underwent lexical semantic change between BCE and CE) and 14 lemmas as "unchanged" (meaning that they did not undergo lexical semantic change). For details on the clustering and the annotation, see Schlechtweg et al. (2020) and Schlechtweg et al. (2021). The SemEval task competition and the subsequent article describing a subset of the systems that took part in it offers the first systematic evaluation of state-of-the-art systems for automatic lexical semantic change detection.

Word embedding models build vector representations of a word for every time slice at hand. For two time intervals t_1 and t_2, we then use a similarity measure (usually, cosine similarity) as a proxy to determine the semantic change between the vectors w_{t_1} and w_{t_2} for a specific word between these time slices:

$$\text{cosine\_similarity}(w_{t_1}, w_{t_2}) = \frac{w_{t_1} \cdot w_{t_2}}{\|w_{t_1}\| \|w_{t_2}\|},$$

where $\| \cdot \|$ denotes the Euclidian norm. A high cosine similarity (e.g., close to 1) means no difference for word w between time slices t_1 and t_2, and a low cosine similarity indicates a high difference.

As our ground truth consists of a binary classification (no-change/change, cf. Section 7), we must transform the cosine similarity value, bounded between -1 and 1, into a decision. While manual thresholding on the cosine is usually applied, recent work (Zhou & Li 2020) shows that determining the threshold in a data-driven way is beneficial. We thus follow prior work on Latin and fit a Gamma distribution of the cosine similarities for all target words between t_0 and t_1, and consider every cosine similarity below the 75-quantile value as the threshold for a change decision.[6]

On the other hand, dynamic Bayesian mixture models, such as SCAN and GASC, are designed to infer the smooth evolution of sense probabilities over time. We adapt these methods to detect sense change points as follows. First, we compute the mean and standard deviations of the posterior sense probabilities over time based on the Gibbs samples obtained during inference. Then, we infer that there has been a significant drop or rise of a sense if its posterior mean probability changes by at least two standard deviations over time. In case of a significant drop we infer that a sense disappeared, and in case of a significant increase we infer that a new sense appeared in the data. If sense probabilities do not change significantly over time, we conclude that no meaning change occurred. Note that, unlike SCAN, GASC outputs a sense probability over time for each genre, and we thus check across all genres whether a significant change of sense probability occurred over time. While we adopt this approach for simplicity, change point analysis has been studied extensively in the context of Gaussian dynamic state space models. We refer to West & Harrison (1997) and Frühwirth-Schnatter (2006) for more sophisticated approaches to detect change points, which also allow for returning a probability distribution over change points.

7 Experiments

7.1 Tasks and baselines

We compared SCAN and GASC to a wide range of baselines on the task of detecting binary change in both Latin and Ancient Greek. Perrone et al. (2019) and Vatri et al. (2019) present a gold standard set created for the purpose of evaluating GASC on Ancient Greek. This set consists of the sense annotation of corpus sentences for three words (*mus* ‘mouse’/‘muscle’/‘mussel’, *harmonia* ‘fastening’/ ‘agreement’/‘musical scale, melody’, *kosmos* ‘order’/‘world’/‘decoration’). These lemmas display a high degree of clear-cut polysemy,[7] especially across genres

[6] We thank Jinan Zhou and Jiaxin Li for providing us with their implementation.

[7] By clear-cut polysemy, we mean that the different senses of a word are not strongly related.

(Liddell et al. 1996, Pollitt 1974), and were chosen as “non-changed” words. We considered two additional lemmas, which display a degree of lexical semantic change in the time period under study, *parabole* ‘comparison’/‘parable’ and *paradeisos* ‘garden’/‘paradise’ (McGillivray et al. 2019). *Paradeisos* is an Avestan loan word that first appeared in Greek in the fifth century BCE to indicate a ‘royal park’ and probably became common after the Macedonian conquest of the Persian empire. This word was chosen by the Greek translators of the Pentateuch to refer to the garden of Eden around the third century BCE (Kyrtatas 2007). The meaning of *parabole*, in turn, specialized from that of ‘comparison’ to that of ‘short moral narrative’ with the New Testament (first century CE). For Latin, we made use of the SemEval task’s gold standard, consisting of 26 “changed” lemmas and 14 “non-changed” lemmas between BCE and CE. We start by visualizing the smooth semantic change inferred by GASC, and then compare the ability of dynamic Bayesian mixture models to detect binary semantic change with the state of the art, both on Latin and Ancient Greek.

7.2 Smooth semantic change

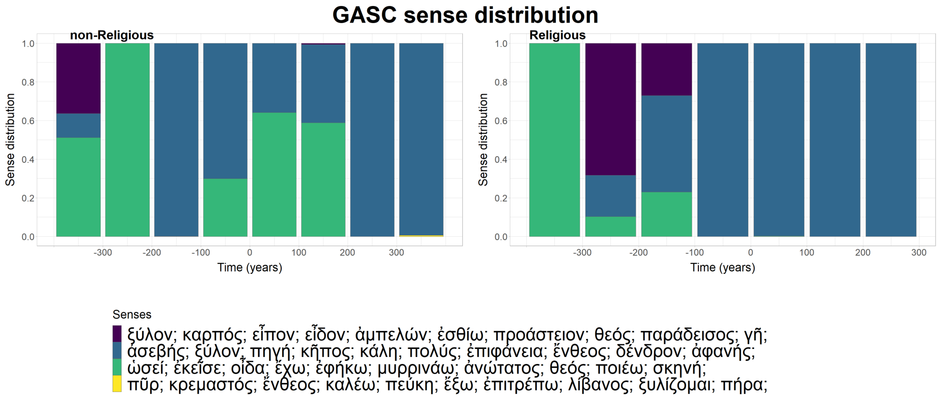


Figure 9.3: Semantic change in Ancient Greek. Visualization of the probability distributions produced by GASC on the Religious genre for the word *paradeisos* (‘garden’/‘paradise’). Negative numbers refer to years BCE.

Dynamic Bayesian mixture models are able to infer the full evolution of sense probabilities over time. In particular, GASC is able to do so for each genre provided as input. Figure 9.3 shows the time distribution of the senses of *paradeisos* outputted by GASC run on the Religious vs. non-Religious genres. The four senses identified by GASC may be interpreted as identifying the meaning ‘garden’ (senses 3-green and 4-yellow), and ‘garden of Eden/(Biblical) paradise’ (senses 1-purple and 2-blue). The two senses are not easily distinguishable (since the Biblical paradise is described as a physical garden) and all senses share a number of words, including, notably, *theos* ‘God’ or the derived adjective *entheos* ‘inspired by God’. However, the first two senses contain a number of words that are easily identifiable as connected to the Biblical narration of the fall of man (e.g. *karpos*, (the forbidden) ‘fruit’ and *esthio*, ‘eat’) while the remaining senses suggest references to other proverbial gardens (e.g. *kremastos* ‘hanging’ garden of Babylon). The diachronic evolution of sense distributions in the plots shows that the Biblical meaning comes to rise around the third century BCE in religious texts, which corresponds precisely to the beginning of the translation of the Bible in Greek, and will prevail throughout the Christian era. The graph displaying the computed distribution of senses in non-religious genres captures well the fact that between the first century BCE and the second century CE *paradeisos* is attested a number of times in the works of historians and geographers represented in the corpus. After the third century, this word is very rarely attested in the works included in the Diorisis corpus and almost half of its occurrences in non-religious texts refer to the Biblical garden of Eden.

7.3 Binary semantic change

Next, we evaluated the ability to recover ground truth about binary semantic change on both Latin and Ancient Greek. For Latin, we recall that ground truth consists of 40 target lemmas, 26 of which underwent semantic change. We ran the genre-aware baselines by specifying whether a text belongs to the Christian genre or not. Results in Table 9.1 show that Bayesian models are highly competitive with the best baseline obtained in the SemEval task, with SCAN achieving the highest F1 score. This is striking as dynamic Bayesian mixture models are designed for capturing smooth semantic change over time, rather than binary semantic change across a pair of time points. In addition, only focusing on non-Christian genres decreases the recall of SGNS and TR. This is expected as the 26 lemmas that underwent semantic change did so due to the rise of a new Christian meaning.

We then evaluated each method on Ancient Greek, further adapting SGNS and TR to use genre information and focus on technical and narrative texts. To evaluate GASC, we use Religious as the genre for *parabole* and *paradeisos*, while Technical and Narrative for *mus*, *harmonia* and *kosmos*, with results being averaged across the five words. Results are shown in Table 9.3. While the small number of target words makes these results mainly illustrative, dynamic Bayesian mixture models emerge as competitive approaches. Consistently with Latin, GASC and SCAN outperform most baselines. To better understand how differently SCAN and TR (the two best-perfoming systems) behave, we refer to the confusion matrix in Table 9.2.

Table 9.1: Semantic change in Latin. Comparison of SCAN and GASC with SGNS, TR and the best baseline from the SemEval task. Results in terms of precision, recall, and F1-score ("F1") averaged across all 40 available words. Results for TR_NAIVE are by Zhou & Li (2020).

Latin (BCE/CE)	Precision	Recall	F1 score
SCAN	0.684	1.000	**0.813**
GASC	0.650	0.920	0.762
SGNS_NOT-christian	1.000	0.308	0.471
SGNS_NAIVE	0.900	0.347	0.500
TR_NOT-christian	0.667	0.231	0.343
TR_NAIVE	0.769	0.769	0.769
Best baseline	0.650	1.000	0.788

Table 9.2: Confusion matrix for binary change in Latin for SCAN and Temporal Referencing. TP = true positive, TN = true negative, FP = false positive, FN = false negative.

System	TP	TN	FP	FN
SCAN	26	2	12	0
TR_NAIVE	20	8	6	6

8 Discussion and conclusion

This work investigates semantic change in Latin and Ancient Greek through several state-of-the-art models. We adapted, discussed and applied a number of algorithms to the case of ancient languages. The adoption of quantitative corpus-based approaches in historical linguistics is growing (Jenset & McGillivray 2017). However, computational approaches to lexical semantic change detection have not yet been widely used in historical linguistics research (McGillivray 2020),

Table 9.3: Semantic change in Ancient Greek. Comparison of SGNS, TR, GASC and SCAN on the task of detecting binary semantic change. Results in terms of precision, recall, and F1-score (“F1”) are averaged across the 5 available words.

Ancient Greek	Precision	Recall	F1 score
GASC	0.600	1.0	**0.750**
SCAN	0.500	0.667	0.571
SGNS_NOT-technical	0.333	0.500	0.400
SGNS_NOT-narrative	0.333	0.500	0.400
SGNS_technical	0.000	0.000	0.000
SGNS_narrative	0.000	0.000	0.000
SGNS_NAIVE	0.333	0.500	0.400
TR_NOT-technical	0.400	1.0	0.571
TR_NOT-narrative	0.333	0.500	0.400
TR_technical	0.000	0.000	0.000
TR_narrative	0.500	1.0	0.667
TR_NAIVE	0.333	0.500	0.400

although a few steps in this direction have been taken (see e.g. Keersmaekers 2020, Rodda et al. 2019, and McGillivray et al. 2019). In spite of their limited use in lexical semantic change detection, dynamic Bayesian mixture models allow practitioners to embed domain expert knowledge and provide interpretable outputs.

We provided a systematic comparison of SCAN and GASC, two recent models from this family, with state-of-the-art embedding-based models, such as SGNS and Temporal Referencing. In addition, we transformed embedding models to account for genre information and provided a new evaluation framework to detect binary semantic change based on expert-annotated data.

Our experiments show that Bayesian models are highly competitive at detecting binary change, beating all baselines on Ancient Greek and Latin. These results, together with the ability to provide full representations of the evolution of word senses, indicate Bayesian dynamic mixture models as successful approaches to study semantic change in ancient languages.

This work can also be seen as a step towards the development of richer evaluation schemes and models that can embed expert judgement. We have shown how including genre can improve the understanding of the historical development of words in a corpus. We argue that the next process to be captured from semantic change models is the archiving of historical texts. The entirety of the relevant documents extant at any time in history is an obvious reference population against which we perform inference. While any analysis based on a currently extant corpus could be biased, Bayesian models embedding historical domain knowledge enable us to de-bias the study (e.g., by accounting for missing texts when inferring the popularity of a sense). There are essentially three different necessary conditions for a text to be extant at any given time. The first is the decision of a librarian to add a particular document to a library, the second is whether or not that text is preserved or destroyed during the passage of time, and the third is the (in)ability of researchers to access documents extant at the current time. A Bayesian analysis enables us to embed a probabilistic description of such a development. For example, many texts within a given corpus will have their own associated provenance, which can be used to help inform the nature of the likely extant corpus. This allows historical insights and extra data to be drawn into the analysis and better inform historical conjectures. The explicit development of such models is ongoing, and we will report our findings in future work.

Acknowledgements

VP's contribution was done prior to joining Amazon. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and the seed funding grant SF042. SH's work is funded by the project *Towards computational lexical semantic change detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184).

Abbreviations

NLP	natural language processing
SGNS	skip-gram with negative sampling

References

Ahmed, Amr & Eric P. Xing. 2012. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. *UAI* abs/1203.3463.

Bamman, David & Gregory Crane. 2011. Measuring historical word sense variation. In *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries*, 1–10. ACM.

Basile, Pierpaolo & Barbara McGillivray. 2018. Exploiting the web for semantic change detection. In *Discovery science* (Lecture Notes in Computer Science 11198). Springer-Verlag.

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65. 487–517.

Blei, David M. & John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning*, 113–120.

Boyd-Graber, Jordan, David Blei & Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 EMNLP-CoNLL*, 1024–1033. http://www.aclweb.org/anthology/D/D07/D07-1109.

Cassotti, Pierluigi, Annalina Caputo, Marco Polignano & Pierpaolo Basile. 2020. GM-CTSC at SemEval-2020 task 1: Gaussian mixtures cross temporal similarity clustering. In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Barcelona: ACL.

Cavallin, Karin. 2012. Automatic extraction of potential examples of semantic change using lexical sets. In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, 370–377. Vienna: ÖGAI.

Clackson, James. 2011. *A companion to the Latin language*. London: Wiley-Blackwell.

Clarke, Michael. 2010. Register Variation. In Egbert J. Bakker (ed.), *A companion to the Ancient Greek language*, 120–33. Chichester/Malden: Wiley-Blackwell.

Cook, Paul, Jey Han Lau, Diana McCarthy & Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014: Technical papers*, 1624–1635. Dublin: ACL. https://www.aclweb.org/anthology/C14-1154.

Davies, Mark. 2002. *The corpus of historical American English (COHA): 400 million words, 1810-2009*. Provo: Brigham Young University.

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. *Corpora* 7(2). 121–157.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019: Volume 1 (Long and short papers)*, 4171–4186. Minneapolis: ACL. DOI: 10.18653/v1/N19-1423.

Dubey, Avinava, Ahmed Hefny, Sinead Williamson & Eric P. Xing. 2013. A non-parametric mixture model for topic modeling over time. *SDM* 2013. 530–538.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of ACL 2019*, 457–470. Florence: ACL. DOI: 10.18653/v1/P19-1044.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Eger, Steffen & Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of ACL 2016 (Volume 2: Short papers)*, 52–58. Berlin: ACL. DOI: 10.18653/v1/P16-2009.

Falk, Ingrid, Delphine Bernhard & Christophe Gérard. 2014. From the culinary to the political meaning of *quenelle*: Using topic models for identifying novel senses (De la quenelle culinaire á la quenelle politique : identification de changements sémantiques á l'aide des Topic Models). In *Proceedings of TALN 2014 (Volume 2: Short papers)*, 525–530. Marseille, France: Association pour le Traitement Automatique des Langues. https://www.aclweb.org/anthology/F14-2023.

Frermann, Lea. 2017. *Bayesian models of category acquisition and meaning development*. Edinburgh: University of Edinburgh. (Doctoral dissertation).

Frermann, Lea & Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the ACL* 4. 31–45. DOI: 10.1162/tacl_a_00081.

Frossard, Esteban, Mickael Coustaty, Antoine Doucet, Adam Jatowt & Simon Hengchen. 2020. Dataset for temporal analysis of English-French cognates. In *Proceedings of LREC 2020*, 855–859. Marseille: ELRA. https://www.aclweb.org/anthology/2020.lrec-1.107.

Frühwirth-Schnatter, S. 2006. *Finite mixture and Markov switching models*. 1st edn. Berlin: Springer.

Gale, William A., Kenneth W. Church & David Yarowsky. 1992. One sense per discourse. In *Speech and natural language: Proceedings of a workshop held at Harriman, New York, February 23–26, 1992*. https://www.aclweb.org/anthology/H92-1045.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Hengchen, Simon. 2017. *When does it mean? Detecting semantic change in historical texts*. Brussels: Université libre de Bruxelles. (Doctoral dissertation).

Hengchen, Simon, Nina Tahmasebi, Dominik Schlechtweg & Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 341–372. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040322.

Jatowt, Adam, Ricardo Campos, Sourav S. Bhowmick, Nina Tahmasebi & Antoine Doucet. 2018. Every word has its history: Interactive exploration and Visualization of word sense evolution. In *Proceedings of CIKM 2018*, 1899–1902. ACM.

Jawahar, Ganesh & Djamé Seddah. 2019. Contextualized diachronic word representations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 35–47. Florence: ACL.

Jenset, Gard B. & Barbara McGillivray. 2017. *Quantitative historical linguistics. A corpus framework*. Oxford: Oxford University Press.

Keersmaekers, Alek. 2020. *A computational approach to the Greek papyri: Developing a corpus to study variation and change in the post-classical Greek complementation system*. KU Leuven. (Doctoral dissertation).

Kessler, Brett, Geoffrey Numberg & Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of EACL 1997*, 32–38. ACL.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore: ACL. DOI: 10.3115/v1/W14-2517.

Korb, K. B. & A. E. Nicholson. 2009. *Bayesian artificial intelligence*. Cleveland, OH: CRC Press.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski & Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of COLING 2018*, 1384–1397. Santa Fe: ACL.

Kyrtatas, Dimitris J. 2007. Paradeisos. In Anastasios-Phoibos Christidis (ed.), *A history of ancient Greek: From the beginnings to Late Antiquity*, 1137–40. Cambridge: Cambridge University Press.

Lau, Jey Han, Paul Cook, Diana McCarthy, Spandana Gella & Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of ACL 2014 (Volume 1: Long papers)*, vol. 1, 259–270. ACL.

Leiwo, Martti, Hilla Halla-aho & Marja Vierros. 2012. *Variation and change in Greek and Latin. Papers and monographs of the Finnish Institute at Athens*. Helsinki: Foundation of the Finnish Institute at Athens.

Liddell, Henry George, Robert Scott, Henry Stuart Jones & Roderick McKenzie. 1996. *A Greek-English lexicon*. 9th ed. Oxford: Oxford University Press.

McGillivray, Barbara. 2020. Semantic analysis of historical texts. In Kristen Schuster & Stuart Dunn (eds.), *Routledge international handbook of research methods in digital humanities*. London: Routledge.

McGillivray, Barbara, Simon Hengchen, Viivi Lähteenoja, Marco Palma & Alessandro Vatri. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities* 34(4). 893–907.

McGillivray, Barbara & Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In Paul Bennett, Martin Durrell, Silke Scheible & Richard J. Whitt (eds.), *New methods in historical corpus linguistics*. Tübingen: Narr.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint 1301.3781*.

Mimno, David, Hanna Wallach & Andrew McCallum. 2008. *Gibbs sampling for logistic normal topic models with Graph-based priors*. Paper presented at the NIPS workshop on analyzing graphs, 2008, Whistler, BC.

Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal & Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21(5). 773–798.

Mitra, Sunny, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee & Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of ACL 2014 (Volume 1: Long papers)*, 1020–1029. Baltimore: ACL. DOI: 10.3115/v1/P14-1096.

Montariol, Syrielle & Alexandre Allauzen. 2020. Étude des variations sémantiques á travers plusieurs dimensions (Studying semantic variations through several dimensions). In *Actes de la 6e conférence conjointe journées d'études sur la parole (JEP, 33e édition), traitement automatique des langues naturelles (TALN, 27e édition), rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues (réCITAL, 22e édition). Volume 2 : Traitement automatique des langues naturelles*, 314–322. Nancy: ATALA et AFCP. https://www.aclweb.org/anthology/2020.jeptalnrecital-taln.31.

O'Hagan, A. & J. Oakley. 2014. *SHELF: The Sheffield elicitation framework*. http://www.tonyohagan.co.uk/shelf/.

Pearl, J. 2009. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Dordrecht: Elsevier.

Perrone, Valerio, Paul A. Jenkins, Dario Spanò & Yee Whye Teh. 2017. Poisson random fields for dynamic feature models. *Journal of Machine Learning Research* 18(127). 1–45.

Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith & Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 56–66. Florence: ACL. DOI: 10.18653/v1/W19-4707.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018: Volume 1 (Long papers)*, 2227–2237. New Orleans: ACL. DOI: 10.18653/v1/N18-1202.

Pollitt, Jerome Jordan. 1974. *The ancient view of Greek art: Criticism, history, and terminology* (Yale publications in the History of Art 26). New Haven: Yale University Press.

Rao, Vinayak & Yee Whye Teh. 2009. Spatial normalized gamma processes. In *Advances in neural information processing systems*, 1554–1562.

Rodda, Martina A., B. McGillivray & P. Probert. 2019. Vector space models of ancient Greek word meaning, and a case study on homer. *Traitement Automatique Des Langues* 60. 63–87.

Rodda, Martina A., Marco S.G. Senaldi & Alessandro Lenci. 2016. Panta rei: Tracking semantic change with distributional semantics in ancient Greek. In Pierpaolo Basile et al. (eds.), *Proceedings of third Italian conference on Computational Linguistics (CLiC-it 2016)* (CEUR Workshop Proceedings 1749). http://ceur-ws.org/Vol-1749/paper46.pdf.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers & Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence*, 487–494.

Rudolph, Maja R. & David M. Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of WWW 2018*, 1003–1011. ACM. DOI: 10.1145/3178876.3185999.

Scarlini, Bianca, Tommaso Pasini & Roberto Navigli. 2020. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. English. In *Proceedings of the 12th language resources and evaluation conference*, 5905–5911. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.723.

Schlechtweg, Dominik, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde & Daniel Hole. 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of CoNLL 2017*, 354–367. Vancouver: ACL.

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici & Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of ACL 2019*, 732–746. Florence: ACL.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of SemEval 2020*, 1–23. Barcelona: ACL. https://www.aclweb.org/anthology/2020.semeval-1.1.

Schlechtweg, Dominik, Sabine Schulte im Walde & Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of NAACL 2018*. ACL.

Schlechtweg, Dominik, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky & Barbara McGillivray. 2021. Dwug: A large resource of diachronic word usage graphs in four languages.

Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale & Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of EMNLP-IJCNLP 2019*, 66–76. Hong Kong: ACL.

Smith, Jim Q. 2010. *Bayesian decision analysis: Principles and practice*. Cambridge New York: Cambridge University Press, Cambridge.

Sprugnoli, Rachele, Marco Passarotti & Giovanni Moretti. 2019. *Vir* is to *moderatus* as *mulier* is to *intemperans*: Lemma embeddings for Latin. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Torino: Accademia University Press. DOI: 10.5281/zenodo.3565572.

Tahmasebi, Nina. 2018. A study on Word2Vec on a historical Swedish newspaper corpus. In *Proceedings of the digital humanities in the Nordic countries, 3rd conference, Helsinki, March 7–9, 2018*. (CEUR Workshop Proceedings 2084). University of Helsinki, Faculty of Arts.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278*.

Tahmasebi, Nina & Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of RANLP 2017*, 741–749. Varna: INCOMA Ltd. DOI: 10.26615/978-954-452-049-6_095.

Tang, Xuri, Weiguang Qu & Xiaohe Chen. 2016. Semantic change computation: A successive approach. *World Wide Web* 19(3). 375–415. DOI: 10.1007/s11280-014-0316-y.

Traugott, Elizabeth Closs & Richard B. Dasher. 2001. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Vatri, Alessandro, Viivi Lähteenoja & Barbara McGillivray. 2019. *Ancient Greek semantic change: Annotated datasets and code [dataset]*. DOI: 10.6084/m9.Figshare.C.4445420.

Vatri, Alessandro & Barbara McGillivray. 2018. The Diorisis Ancient Greek corpus. *Research Data Journal for the Humanities and Social Sciences* 3(1). 55–65. DOI: 10.1163/24523666-01000013.

West, Mike & Jeff Harrison. 1997. *Bayesian forecasting and dynamic models.* 2nd edn. Berlin: Springer-Verlag.

Wijaya, Derry Tanti & Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of DETECT 2011*, 35–40. ACM.

Wilkerson, Rachel L. & Jim Q. Smith. 2018. Customised structural elicitation. *arXiv preprint arXiv:1807.03693.*

Zamora-Reina, Frank D. & Felipe Bravo-Marquez. 2020. DCC-Uchile at SemEval-2020 Task 1: Temporal referencing word embeddings. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona: ACL.

Zampieri, Marcos, Shervin Malmasi & Mark Dras. 2016. Modeling language change in historical corpora: The case of Portuguese. In Nicoletta Calzolari et al. (eds.), *Proceedings of LREC 2016.* Portorož: ELRA.

Zhou, Jinan & Jiaxin Li. 2020. TemporalTeller at SemEval-2020 Task 1: Unsupervised lexical semantic change detection with temporal reference. In *Proceedings of the 14th International Workshop on Semantic Evaluation.* Barcelona: ACL.

Chapter 10

Computational approaches to lexical semantic change: Visualization systems and novel applications

Adam Jatowt[a], Nina Tahmasebi[b] & Lars Borin[b]

[a]University of Innsbruck [b]University of Gothenburg

The purpose of this chapter is to survey visualization and user interface solutions for understanding lexical semantic change as well as to survey a number of applications of techniques developed in computational analysis of lexical semantic change. We first overview approaches aiming to develop systems that support understanding semantic change in an interactive and visual way. It is generally accepted that computational techniques developed for analyzing and uncovering semantic change are beneficial to linguists, historians, sociologists, and practitioners in numerous related fields, especially within the humanities. However, quite a few non-professional users are equally interested in the histories of words. Developing interactive, visual, engaging, and easy-to-understand systems can help them to acquire relevant knowledge.

Second, we believe that other fields could benefit from the research outcomes of computational approaches to lexical semantic change. In general, properly representing the meaning of terms used in the past should be important for a range of natural language processing, information retrieval and other tasks that operate on old texts. In the latter part of the chapter, we then focus on current and potential applications related to computer and information science with the underlying question: "How can modeling semantic change benefit wider downstream applications in these disciplines?"

Adam Jatowt, Nina Tahmasebi & Lars Borin. 2021. Computational approaches to lexical semantic change: Visualization systems and novel applications. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 311–339. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040320

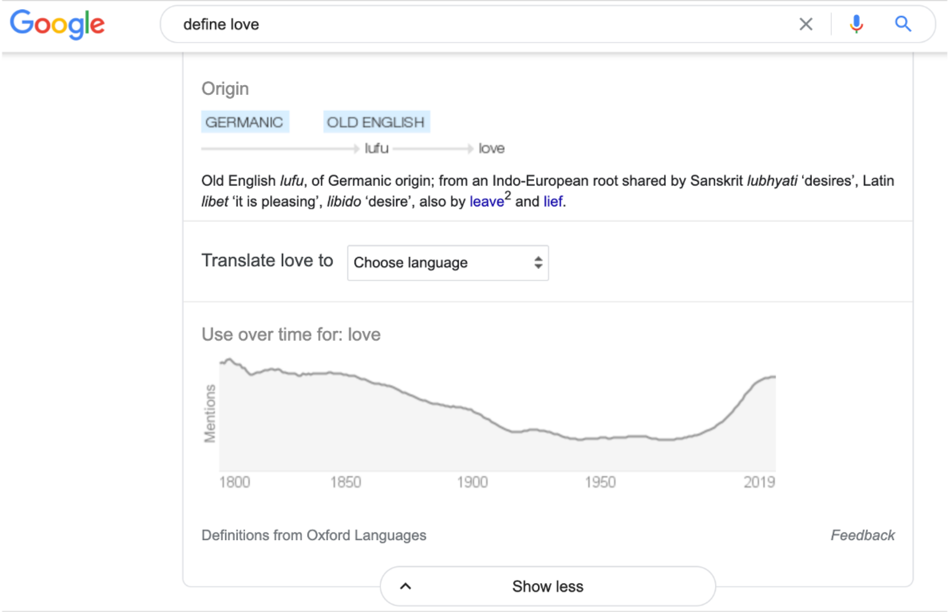
1 Visualization systems supporting manual analysis

It is already evident that computational approaches greatly speed up the research and the resulting discoveries related to semantic change and lexical replacement. The applications of this research progress still remain to be seen, both inside and outside the academic realm. Providing effective methods for detecting changes, their characteristics, timing, and causal factors are all important for our understanding of languages and their evolution. Computational methods support the acquisition of such knowledge and the formation and validation of various kinds of hypotheses. However, other uses of the technology outside the academic field are less discussed. Word meaning change is not only interesting to professionals (e.g., linguists, historians, or librarians) but also to the wider public. For example, many books discussing the origins and evolution of word meaning have been published aiming at a wider readership, suggesting significant interest by average users in the histories of words. We think that computational approaches and especially online interactive systems are important to help to further disseminate knowledge about etymology.

Visual analytics have been increasingly applied in historical linguistics (Schätzle & Butt 2020) by combining automated algorithms with interactive visual components to let us perform effective investigations through data manipulation and presentation. In this context, several online visualization systems and demonstrations supporting manual analysis have been proposed to complement the research methods developed for detecting diachronic conceptual change. They allow for verification of the results obtained from automatic methods or provide novel means for supporting manual determination of diachronic conceptual change and its characteristics. In these systems the level of interactivity and user freedom in querying the data, as well as the provision of features enabling multidimensional analysis play crucial roles. Evaluation of these systems focuses on usability criteria and their user interfaces. Visualization systems tend to be attractive as they provide either a complement to automatic analysis or serve as the main tool for analysis, and not only for professionals and scientists. Many of them are also particularly suited to lay users, especially if the systems are intuitive and highly usable. Below, we discuss representative systems designed for learning about semantic change and we highlight several new directions based on their analysis. Due to copyright restrictions, we can only show the screenshots of few, selected examples.

1.1 Description- and frequency-based approaches

Starting with a simple example, the enhancement of definitions generated by the Google word search engine for "definition queries" is an effective way to disseminate basic information on words' origins and patterns of popularity over time. When given an input word, a standard definition can be complemented by an additional click with a brief textual description of the word's origin as well as its frequency plot over time (see Figure 10.1 for an example). This service provides historical context for word definitions that can support better understanding of queried words and also trigger user interest in word histories by appending complementary basic knowledge about the temporal evolution of word meaning. Although users can see the change in the frequency of a word over time and read brief information on the word's origin, they still essentially have to guess about the word's meaning change over the entire span of time. Nevertheless, this application is worth mentioning because, thanks to the popularity of Google search services, information on word change (albeit rather superficial) can be widely disseminated to the public through this service.



Old English *lufu*, of Germanic origin; from an Indo-European root shared by Sanskrit *lubhyati* 'desires', Latin *libet* 'it is pleasing', *libido* 'desire', also by leave[2] and lief.

Figure 10.1: Snapshot of an example output from Google search engine for the query `define love` where simple information on the origin and popularity of word *love* over time is given (image captured on 2021-01-24).

The *Online Etymology Dictionary*[1] is an easy-to-use system that draws on the etymology of numerous words compiled from several dictionaries. The service returns a short etymological description of an input word with dates indicating the earliest year for which there is a surviving written record of its use.

The *Google Books Ngram Viewer*[2] (Michel et al. 2011) – another notable contribution from the search engine company – is a powerful online application for observing and analyzing the frequency of words or Ngrams over time. It is based on the Google Books Ngrams datasets. The service has frequently been used for digital humanities research and the like (e.g. Michel et al. 2011, Acerbi et al. 2013, Bentley et al. 2014, Pechenick et al. 2015, Iliev et al. 2016). The temporal frequency plots of several words or Ngrams can be contrasted with each other. Users can choose a wildcard search (by putting * in place of a word in a given phrase to obtain the top ten substitutions) or do a case-insensitive search. Analysis based on parts of speech (POS) tags is possible (e.g., plotting frequencies of *tackle* as either a verb or as a noun). It is also possible to plot the frequency based on five composition operators (e.g., summing or subtracting the frequencies of several expressions). Inflection-oriented search can be done (e.g., searching with `book_INF a hotel` returns results for *book, booked, books*, and *booking a hotel*). The Ngram viewer allows the identification of words at the start or end of sentences to be plotted. It provides dependency relations using the `=>` operator. For instance, to understand how often *tasty* was used to modify the word *dessert* one would input (`tasty => dessert`). This search combines frequencies of all instances in which the word *tasty* modifies *dessert*, including *tasty frozen dessert*, and *tasty yet expensive dessert*. Dependencies can be further combined with wildcards (e.g., `drink => *_NOUN` to track frequencies of expressions containing different kinds of beverages as nouns). Nevertheless, because the viewer is mainly based on the frequency signals of words (i.e., probabilities of seeing a given Ngram or a set or composite of Ngrams in a given year), it does not provide a direct means for portraying exactly how a term was used in the past or when its meaning transitions occurred. The viewer is thus best suited to culturomics or cultural text mining studies and is similar to other tools available for general purpose interactive exploration of diachronic corpora (e.g. Michel et al. 2011, Odijk et al. 2014, van Eijnatten et al. 2014, Jatowt & Bron 2016). Despite this limitation, this application is worth mentioning because it uses extremely large datasets coupled with basic manipulation capabilities, even though it does not explicitly support analysis of semantic change.

[1] https://www.etymonline.com/

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

The online interfaces of several diachronic corpora created by Mark Davies at Brigham Young University[3] provide effective options for users. Users can perform simple analyses without the need to write any code. For example, they can generate frequency plots over time, examples of keywords in context (KWIC) at different time points, and listings of collocates. However, the amount and level of detail of the displayed data make it rather difficult to draw broader conclusions about the semantics of words from a longitudinal perspective.

Hilpert & Gries (2008) apply a variant of hierarchical clustering called variability-based neighbor clustering. The idea is to cluster adjacent time units (hence the name "neighbor clustering") if the frequency of a target term does not change much. The resulting dendrogram allows for visual identification of time points of large frequency change, which may indicate increased possibility of diachronic sense shifts (e.g., due to sudden triggers like large events). No context is used for a target word because the method relies only on the the frequency information of a query word, which limits the applicability of this approach in representing diachronic conceptual change of words.

Odijk et al. (2014) demonstrate an interactive environment that visualizes information on the volumes and correlations of words and documents across time. Similar to Michel et al. (2011), their focus is more on understanding historical and social aspects than on shifts in word meaning.

1.2 Context-based semantic approaches

Rohrdantz et al. (2011) use latent Dirichlet application (LDA) to represent the different senses of words and track their intensity of change over time. Twenty-five words before and twenty-five words after the target word are used as the context of the term, following the suggestion given by Schütze (1998) for automatic sense discrimination. This approach allows one to notice various kinds of semantic change in words, such as the broadening or narrowing of senses and the first occurrences of senses, especially as all the topics are shown over time in a single view. According to Rohrdantz et al. (2011), their interactive visualization approach provides the possibility of detecting key patterns at a glance, while at the same time observing the details of the data by zooming in on the occurrences of particular words in their contexts. Additionally, the results of the pairwise comparisons of word senses with respect to their shared contexts are also displayed. The authors, however, restrict their system to only a short time period, demonstrating results on the New York Times Annotated corpus, which spans roughly two decades.

[3] https://corpus.byu.edu/

Heylen et al. (2012) propose using a multidimensional scaling (Cox & Cox 2008) technique with a window length of 4 words before and after the target word and pointwise mutual information for weighting context terms. Heylen et al. (2012) took this approach because they had observed that earlier automatic approaches which use distributional models use them in an indirect, black-box fashion, failing to indicate particular semantic properties and relations that play key roles. Motion charts from the Google Chart tools are then used to visualize occurrences of nouns in a 2D representation of their semantic distances. Hovering a mouse pointer over the bubbles denoting nouns shows the text in which each noun occurs so that users can interpret the precise meaning of the occurrence of the noun. In their case study, the authors focus on Dutch words extracted from Dutch newspaper articles published between 1999 and 2005, which were organized in 218 synsets containing 476 nouns in total. Although they do not use the motion feature of the charts, the authors note that it should be possible to track the centroid of the tokens of a target word over time in the semantic space, and also to show the dispersion of the tokens around the centroid.

Hilpert & Perek (2015) use animations in the form of animated scatterplots to portray change in patterns over time using the metaphor of a petri dish. The authors focused on a single pattern, "many a [noun]" as a case study. Spots on the graphs represent nouns involved in the same pattern, and are plotted next to each other if they have high similarity. The size of the spot is linked to the frequency of a noun or noun type in a particular time unit. During the animation, the changes in the size and distances of spots provide knowledge of different uses of the pattern over time.

Dimensionality reduction techniques such as principal components analysis (PCA), latent semantic analysis (LSA) or the popular t-distributed stochastic neighbor embedding (t-SNE, van der Maaten & Hinton 2008)[4] have been frequently used to plot "trajectories" of word meaning over time in vector spaces using 2D plots. By showing points that represent the meaning of the same words at different years or decades on the same 2D plot (see, e.g., Hamilton et al. 2016a, Kulkarni et al. 2015), and optionally connecting them with arrows, a single static view can show how the words changed their meaning over time, by simply following their "trajectories". Typically some background reference terms are added along these "trajectories" to ground and explain the meaning.

Martinez-Ortiz et al. (2016) introduce a system called *ShiCo* for visualizing shifting concepts of Dutch words over time. It measures change in words used to refer to concepts based on a model previously introduced by Kenter et al. (2015).

[4] https://lvdmaaten.github.io/tsne/

This model requires a series of semantic spaces that are constructed by training word embeddings (e.g., word2vec) for different units of time (typically, each unit spans 10 years). It is based on two steps, generation and aggregation. The generation step works in an iterative fashion. An initial seed set is selected that typically consists of a small number of user-provided terms. Then, words most semantically similar to the seed set are found based on similarity values between word embeddings. A semantic graph is constructed from these terms, and the central terms are extracted using graph centrality measures. Next, the central terms are used as the seed set for the next iteration of the generation step. In the aggregation step, the lists of words produced in the generation step are aggregated to generate the final word lists to be presented to the user. The visualization by Martinez-Ortiz et al. (2016) is composed of two kinds of complementary graphs: a stream graph and a series of network graphs. The former shows color-coded streams for each term; the stream sizes represent the relative importance of the term in a period. This importance is measured either as a term count in each time unit or as a sum of the similarities of the term to the seed terms. The network graphs for each time unit display the relations between terms in the time unit.

Xu & Crestani (2017) use term clouds and a heatmap to visualize semantic shifts of target words by utilizing sequentially trained embedding vectors with initialization based on previous time periods, as proposed by Kim et al. (2014). They use *The New York Times* and *National Geographic Magazine* articles for the underlying datasets, which span about 110 years. Following Martinez-Ortiz et al. (2016), a temporal semantic similarity word cloud is used to show terms most similar to a target query for a given time unit. As in standard term clouds, the font size of the terms is linked to their similarity to the query word. Heatmap views let users see the similarity values of the terms most similar to the target term in each year, using colors. The y-axis of the heatmap is a list of words and the x-axis is a list of temporal periods such that for each given word (each row) one can understand the pattern of the change in the similarity of this term to the target term (also called anchor term). The results from the *The New York Times* and *National Geographic Magazine* are then contrasted with each other.

Jatowt & Duh (2014) describe an analytical framework that incorporates different types of similarity plotting. The plots include across-time self-similarity, a decade-to-decade similarity heatmap, across-time sentiment analysis, diachronic comparative word analysis, and key context term listing. They use both the Google Books Ngrams and COHA datasets. The signals from the different views (e.g. frequency analysis, semantic analysis, and sentiment analysis) can be combined to allow for multi-evidence-based reasoning on diachronic conceptual change. Two different word representations are used: a simple bag-of-words, and a distance-aware bag-of-words, where the distance used is the relative position of a context term to the target term. For simplicity, the sentiment values of context terms are assumed to remain stable over time when computing the change of the sentiment of the target term's context.

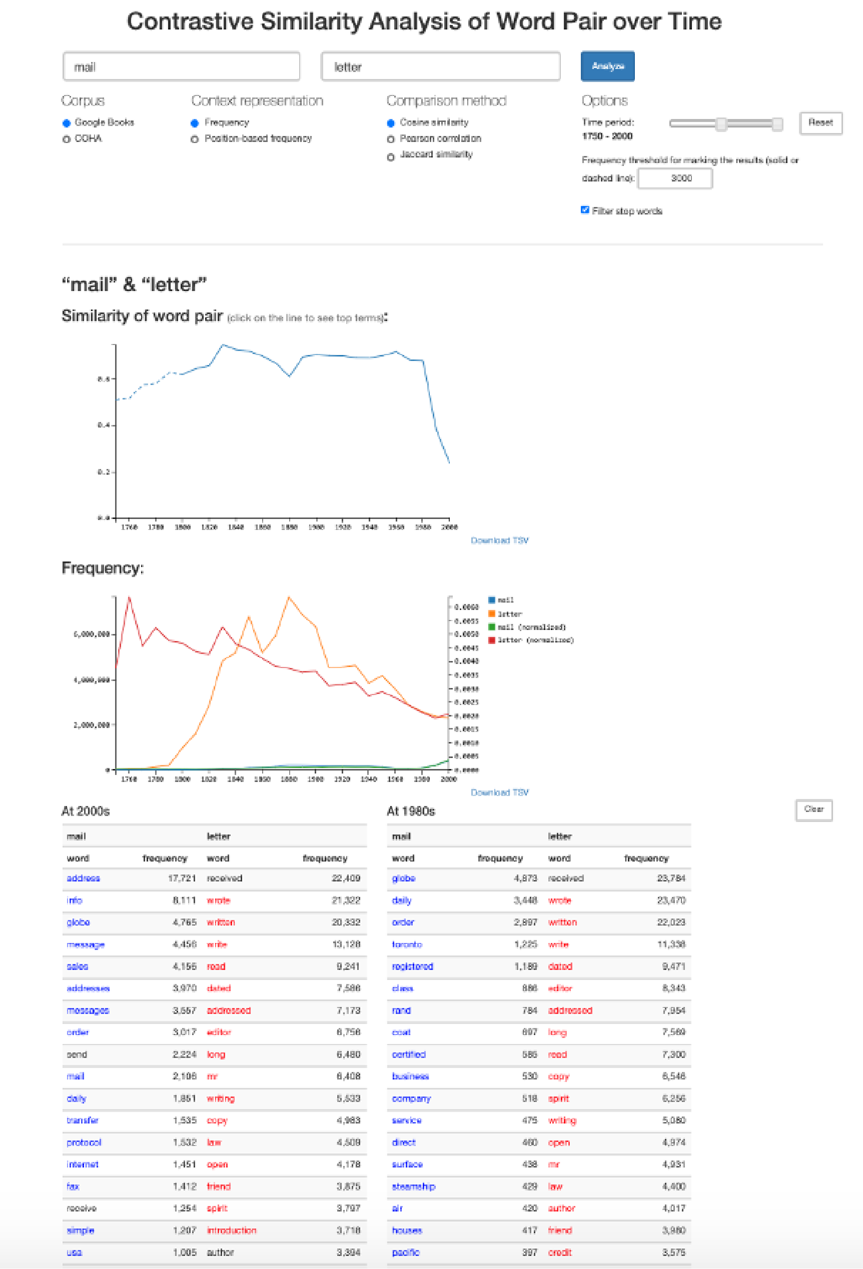
Their work resulted in the development of an online interactive system for diachronic conceptual change analysis (Jatowt et al. 2018) (see Figures 10.2, 10.3 and 10.4).[5] The system enables detailed analysis of diachronic conceptual change from different viewpoints, such as word's context comparison across-time, temporal term cloud, and temporal term tree generation. It can also perform contrastive analysis of word pairs (see Figure 10.2) of larger groups of words such as synonyms. The results can be generated using both the Google Books Ngrams and COHA datasets. Pearson correlation, cosine similarity, and Jaccard similarity can be used as similarity measures of word representations from different time points. Jatowt et al. (2018) recommend that diachronic conceptual change over time should be contrasted with term frequency plots (as also suggested by Kim et al. 2014), since together both provide a more informed view on how often and in what sense a term was used in the past. Any conclusions drawn from semantic change plots should be treated with caution when the frequency plot of a target term shows a low utilization rate. To better visualize the word change in relation to the average change of other words, the degree of the target word's change over time is also displayed with reference to the average change of words in the same frequency bin as the target word.

The system has another novel feature that allows the change in individual context terms over time to be investigated in the form of a *time-enhanced term cloud* (see Figure 10.3) and *time-enhanced term tree* (see Figure 10.4). Finally, the framework provides a unique functionality to track the semantic shifts of entire concepts represented as word sets, for example, the concept of a vehicle represented by words like *auto*, *automobile*, *car*, *truck*, and so on.

Hellrich et al. (2018) proposed JeSemE, the Jena Semantic Explorer,[6] which is an interactive website for visually exploring temporal information on word meanings and lexical emotions on the basis of five large diachronic text corpora in English and German, including COHA and Google Books English fiction. A unique feature of this system is the provision of predicted emotion values of words over time based on the valence-arousal-dominance (VAD) scheme of Bradley & Lang 1994. It also shows similar words and specific contexts of a query word. Figure 10.5 and 10.6 show example outputs.

[5] http://www.okayama.silk.jp/WordEvolution/

[6] http://jeseme.org/



Contrastive Similarity Analysis of Word Pair over Time

mail

letter

Corpus

● Google Books
○ COHA

Context representation

● Frequency
○ Position-based frequency

Comparison method

● Cosine similarity
○ Pearson correlation
○ Jaccard similarity

Options

Time period: 1750 - 2000

Frequency threshold for marking the results (solid or dashed line): 3000

☑ Filter stop words

“mail” & “letter”

Similarity of word pair (click on the line to see top terms):

Frequency:

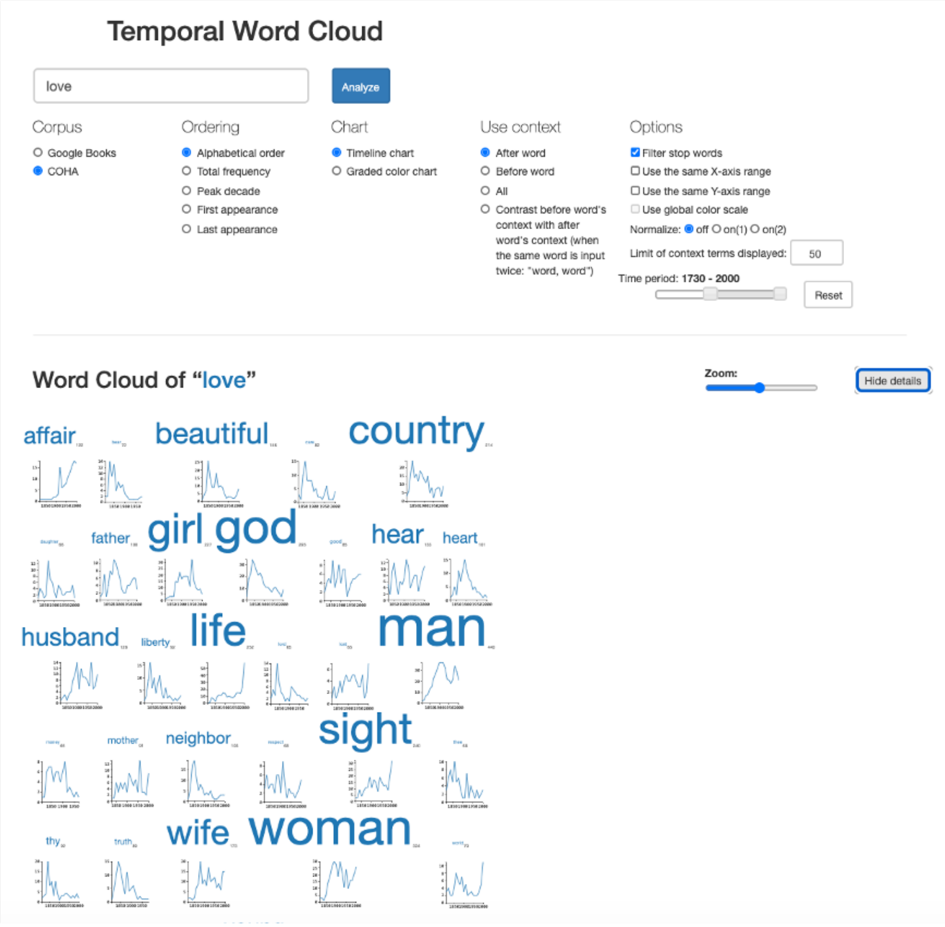
At 2000s

mail		letter	
word	frequency	word	frequency
address	17,721	received	22,409
info	8,111	wrote	21,322
globe	4,765	written	20,332
message	4,456	write	13,128
sales	4,156	read	9,241
addresses	3,970	dated	7,586
messages	3,557	addressed	7,173
order	3,017	editor	6,758
send	2,224	long	6,480
mail	2,106	mr	6,408
daily	1,851	writing	5,533
transfer	1,535	copy	4,983
protocol	1,532	law	4,509
internet	1,451	open	4,178
fax	1,412	friend	3,875
receive	1,254	spirit	3,797
simple	1,207	introduction	3,718
usa	1,005	author	3,394

At 1980s

mail		letter	
word	frequency	word	frequency
globe	4,873	received	23,784
daily	3,448	wrote	23,470
order	2,897	written	22,023
toronto	1,225	write	11,338
registered	1,189	dated	9,471
class	888	editor	8,343
rand	784	addressed	7,954
coat	697	long	7,569
certified	585	read	7,300
business	530	copy	6,548
company	518	spirit	6,256
service	475	writing	5,080
direct	460	open	4,974
surface	438	mr	4,931
steamship	429	law	4,400
air	420	author	4,017
houses	417	friend	3,980
pacific	397	credit	3,575

Figure 10.2: Snapshot of an example output from the diachronic conceptual change analysis system for comparing the words *mail* and *letter* that displays their similarity plot over time, frequency plots and the contrasted lists of top-frequent context term at 1980s and 2000s (image captured on 2021-01-24).



Temporal Word Cloud

love

Analyze

Corpus

○ Google Books

● COHA

Ordering

● Alphabetical order

○ Total frequency

○ Peak decade

○ First appearance

○ Last appearance

Chart

● Timeline chart

○ Graded color chart

Use context

● After word

○ Before word

○ All

○ Contrast before word's context with after word's context (when the same word is input twice: "word, word")

Options

☑ Filter stop words

□ Use the same X-axis range

□ Use the same Y-axis range

□ Use global color scale

Normalize: ● off ○ on(1) ○ on(2)

Limit of context terms displayed: 50

Time period: 1730 - 2000

Reset

Word Cloud of "love"

Zoom:

Hide details

Figure 10.3: Snapshot of an example output from the diachronic conceptual change analysis system which shows a part of temporal term cloud of context words for the input word *love* (image captured on 2021-01-24).

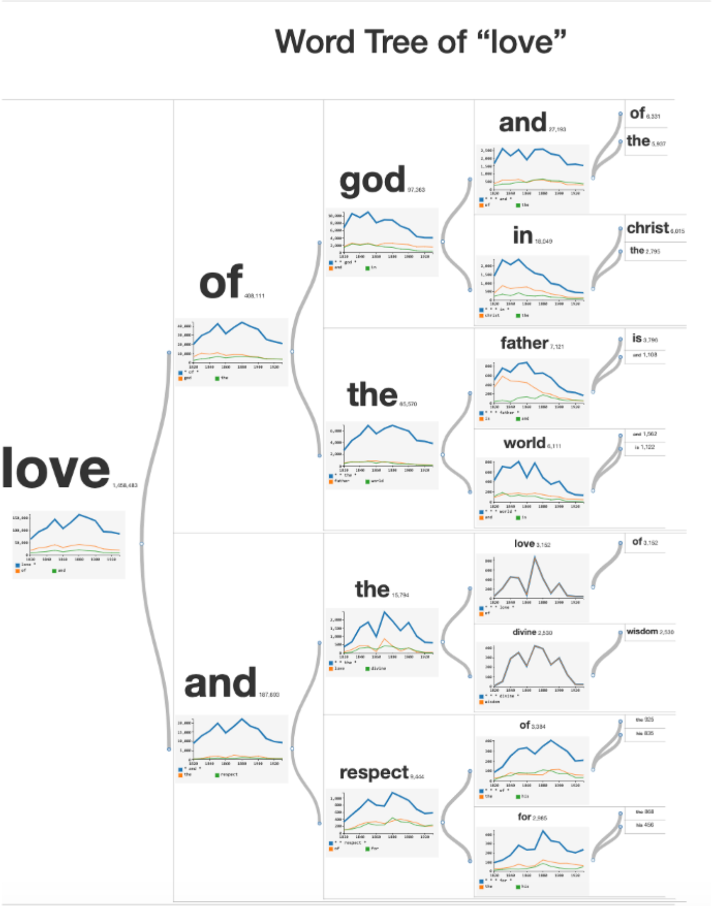
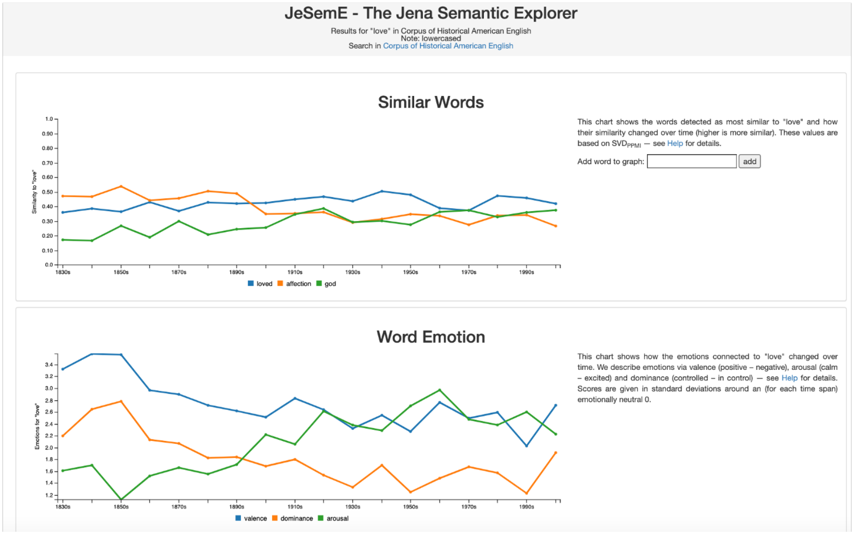


Figure 10.4: Snapshot of an example output from the diachronic conceptual change analysis system which shows a part of temporal term tree for the input word *love* (image captured on 2021-01-24).



JeSemE - The Jena Semantic Explorer

Results for "love" in Corpus of Historical American English
Note: lowercased
Search in Corpus of Historical American English

Similar Words

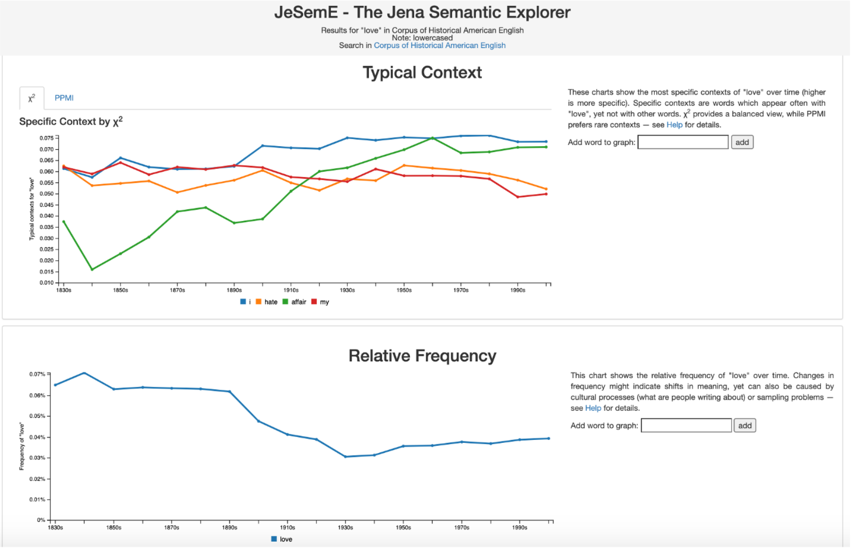
This chart shows the words detected as most similar to "love" and how their similarity changed over time (higher is more similar). These values are based on SVD_{PPMI} — see Help for details.

Add word to graph: add

Word Emotion

This chart shows how the emotions connected to "love" changed over time. We describe emotions via valence (positive – negative), arousal (calm – excited) and dominance (controlled – in control) — see Help for details. Scores are given in standard deviations around an (for each time span) emotionally neutral 0.

Figure 10.5: Snapshot of an example output from JeSemE for the word *love* which shows the similarity of similar words to the input word and associated emotions over time (image captured on 2021-01-24).



JeSemE - The Jena Semantic Explorer

Results for "love" in Corpus of Historical American English
Note: lowercased
Search in Corpus of Historical American English

Typical Context

χ^2 PPMI

Specific Context by χ^2

These charts show the most specific contexts of "love" over time (higher is more specific). Specific contexts are words which appear often with "love", yet not with other words. χ^2 provides a balanced view, while PPMI prefers rare contexts — see Help for details.

Add word to graph: add

Relative Frequency

This chart shows the relative frequency of "love" over time. Changes in frequency might indicate shifts in meaning, yet can also be caused by cultural processes (what are people writing about) or sampling problems — see Help for details.

Add word to graph: add

Figure 10.6: Snapshot of an example output from JeSemE for the word *love* which shows specific context words and relative word frequency over time (image captured on 2021-01-24).

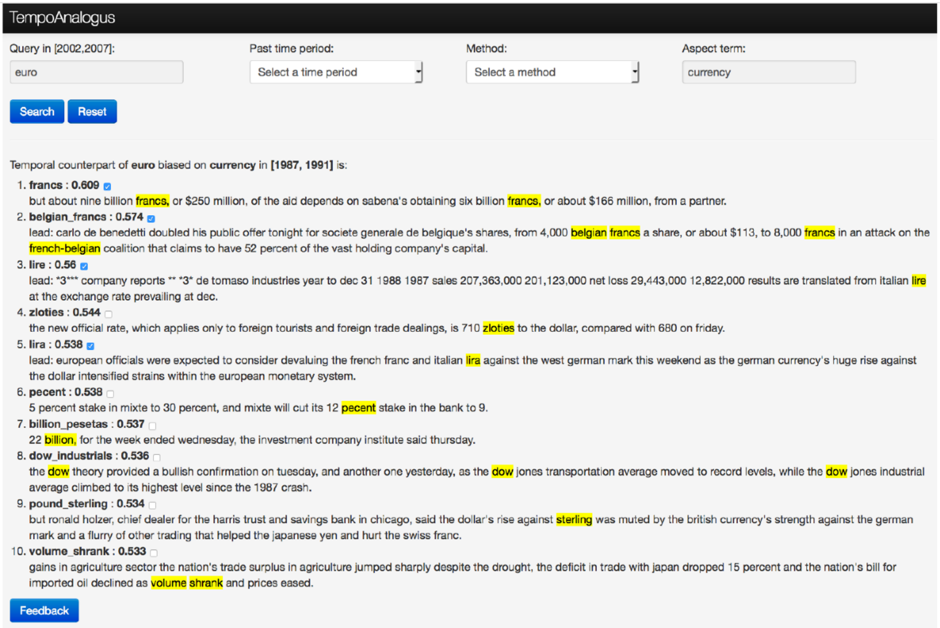
1.3 Dictionary-based approaches

Theron & Fontanillo (2015) demonstrate an interactive visual tool for advanced analysis of the data in Spanish historical dictionaries. Their approach is unique as they utilize different editions of Spanish language dictionaries over time: the 1780, 1817, 1884, 1925, 1992 and 2001 editions provided by the Royal Spanish Academy. In this method the dictionary editions are arranged in a matrix in columns (right to left in chronological order), while the meanings of a word are placed on the rows (top to bottom in ascending order). Lines are drawn to connect related meanings across time, with the connection computed using NIST or BLEU metrics (Zhang et al. 2004), which are frequently utilized in evaluating machine translation or summarization accuracy. Starting from the most recent dictionary, a particular meaning is connected to its closest meaning in the previous dictionary; if there is nothing that satisfies the predefined similarity threshold, then the procedure is repeated with the older dictionary. Connecting lines can have branches in cases of bifurcation or merging of meanings. Theron & Fontanillo (2015) call the resulting diagrams *diachronlex diagrams*. The diagrams can be further improved by collapsing nearby lines with similar temporal patterns or by simplifying branches. Furthermore, users with editing rights can annotate meanings or change their associations.

1.4 Systems for analyzing lexical replacement

Mazeika et al. (2011) focus on semantically similar entities from different time periods to provide visual support in analyzing lexical replacement, particularly for entities. They extracted named entities from the Yet Another Great Anthology (YAGO) database to provide a visual analytics tool to analyze the evolution of named entities in the New York Times Annotated Corpus. Name changes are not tracked, but the tool offers a visualization of the evolution of an entity in relation to other entities.

Investigation of lexical replacement and temporal analogy are also possible in the aspect-based temporal analog retrieval (ATAR) system (Zhang et al. 2019) which uses perceptrons to compute transformations between present and past vector spaces trained on the present and past participles, respectively. For a user query (e.g., *euro*) and its defining aspect or sense (e.g., *currency*) a list of analogical terms is produced based on the analysis of its past document collection, together with an extracted representative sentence for each output term (see Figure 10.7). The sentence provides a typical context in which an analogous word was used in the past, using the output style dubbed KWECT (keyword in exemplar context at time), similar to the traditional KWIC (keyword in context) style.



Query in [2002,2007]: euro

Past time period: Select a time period

Method: Select a method

Aspect term: currency

Temporal counterpart of **euro** biased on **currency** in **[1987, 1991]** is:

1. **francs : 0.609** ☑
   but about nine billion francs, or $250 million, of the aid depends on sabena's obtaining six billion francs, or about $166 million, from a partner.
2. **belgian_francs : 0.574** ☑
   lead: carlo de benedetti doubled his public offer tonight for societe generale de belgique's shares, from 4,000 belgian francs a share, or about $113, to 8,000 francs in an attack on the french-belgian coalition that claims to have 52 percent of the vast holding company's capital.
3. **lire : 0.56** ☑
   lead: *3*** company reports ** *3* de tomaso industries year to dec 31 1988 1987 sales 207,363,000 201,123,000 net loss 29,443,000 12,822,000 results are translated from italian lire at the exchange rate prevailing at dec.
4. **zloties : 0.544** ☐
   the new official rate, which applies only to foreign tourists and foreign trade dealings, is 710 zloties to the dollar, compared with 680 on friday.
5. **lira : 0.538** ☑
   lead: european officials were expected to consider devaluing the french franc and italian lira against the west german mark this weekend as the german currency's huge rise against the dollar intensified strains within the european monetary system.
6. **pecent : 0.538** ☐
   5 percent stake in mixte to 30 percent, and mixte will cut its 12 pecent stake in the bank to 9.
7. **billion_pesetas : 0.537** ☐
   22 billion, for the week ended wednesday, the investment company institute said thursday.
8. **dow_industrials : 0.536** ☐
   the dow theory provided a bullish confirmation on tuesday, and another one yesterday, as the dow jones transportation average moved to record levels, while the dow jones industrial average climbed to its highest level since the 1987 crash.
9. **pound_sterling : 0.534** ☐
   but ronald holzer, chief dealer for the harris trust and savings bank in chicago, said the dollar's rise against sterling was muted by the british currency's strength against the german mark and a flurry of other trading that helped the japanese yen and hurt the swiss franc.
10. **volume_shrank : 0.533** ☐
   gains in agriculture sector the nation's trade surplus in agriculture jumped sharply despite the drought, the deficit in trade with japan dropped 15 percent and the nation's bill for imported oil declined as volume shrank and prices eased.

Figure 10.7: Snapshot of an example output from aspect-based temporal analog retrieval system (ATAR) for the query word *euro* under the aspect of currency (image captured on 2021-01-24).

1.5 Summary and observations

Visualization and analysis of diachronic conceptual change belong to an emerging and powerful research field of interactive visualization for computational linguistics (Collins et al. 2008). Its purpose is to let users understand models of language and their abstract representations, and to visually uncover patterns in language. In view of the inherent complexity of tracking word senses and understanding their shifts over time, we expect an increase in the availability and popularity of visual, interactive approaches to diachronic corpora. A similar conclusion was reached by Tang (2018), who considers further use of data visualization techniques to prove hypotheses as one of the core issues to be solved. Below we list several approaches and future directions.

- Easy to use and attractive services should be built to allow non-professional users to freely investigate histories of any words they are interested in, and to appreciate their language. Entertaining visualizations and explanations would attract many interested visitors. Automatically generating accessible explanations and suggesting interesting words to explore would be beneficial (see Section 2.4 for the discussion of example query recommendation techniques).

- It is often difficult to precisely determine the exact time of sense change, let alone accurately determine the nature of the change. Sometimes several conflicting conclusions or hypotheses can appear simultaneously valid, prompting scientists and professionals to look at the results from different angles and use different datasets as well as visualization techniques. Hence, frameworks providing multiple views or analysis angles, and using parallel datasets, should be especially useful (e.g. Jatowt et al. 2018, Kalouli et al. 2019, Hellrich et al. 2018). Related to this is the role of a word's frequency over time as calculated for the particular corpus used in the analysis. This can work as a confidence measure for observed semantic change. It is is also helpful to contrast semantic analysis results of similar or related words, or words associated with the same concept (e.g. Jatowt et al. 2018).

- Previous analyses (Hamilton et al. 2016b, Pagel et al. 2007, Lieberman et al. 2007) have revealed differing average degrees of change of very frequent and less frequent words, although the influence of frequency on the semantic change degree was later found to be smaller than expected (Dubossarsky et al. 2017). Investigating historical data such as the degree of the target word's change over time could be referenced to the average degree of change of words in the same frequency bin as the bin of the target word.[7]

- When evaluating these systems, it is common practice to investigate particular cases to determine whether the results support expectations or existing knowledge about diachronic conceptual change. The same type of evaluation is done with general purpose information visualization systems (Carpendale 2008). The investigation can extend to checking whether novel kinds of information can be obtained. We think that more systematic and extensive evaluation frameworks should be applied to determine whether new systems really help to find changes. New systems should also be compared with other systems to determine their strengths as well as their weaknesses. The unsupervised lexical semantic change detection task (Schlechtweg et al. 2020) at SemEval-2020 is an example of a standardization initiative to evaluate algorithms based on a shared dataset and evaluation metrics.

[7] As, for example, in Jatowt et al. (2018).

- Finally, systems designed primarily for visualization and interactive analysis of syntactic change in historical linguistics such as HistoBankVis (Schätzle & Butt 2020) and ParHistVis (Kalouli et al. 2019) can provide novel insights for building interactive tools for semantic analysis. Comprehensive visual analytics frameworks for historical linguistics could embrace both semantic and syntactic change and their interrelation to enable the comprehensive study of change in linguistic phenomena over time.

2 Applications of computational analysis of semantic change and lexical replacement

The remainder of this chapter deals with several applications of approaches designed for computational modeling, analysis of semantic change, and lexical replacement. In particular, we focus on the use of the technologies outside the core objective of analyzing the change in word meaning per se, that is, aiming to reveal knowledge of a word's history. The techniques developed for diachronic conceptual analysis and the findings from their use can be beneficial for various applications and services that deal with old texts or documents in long-term document archives. We discuss some current as well as promising future applications, especially within computer and information sciences. We note that our overview is in no way definitive and exhaustive, as many computational processes applied to old texts could benefit from the techniques and discoveries in the field of computational approaches to semantic change.

2.1 Semantics-aware culturomics

Michel et al. (2011) coined the term *culturomics* – the study of cultural and historical phenomena based on large textual data. In their seminal paper the authors demonstrate changes in the frequencies of selected words that reveal high-level cultural or abstract change occurring in a society over time. As one example, they contrast the popularity plots of the words *men* and *women* to provide evidence for the increasing social role and emancipation of women in recent decades. The work by Michel et al. (2011) inspired many similar studies using the Google Books Ngrams datasets (e.g. Acerbi et al. 2013, Bentley et al. 2014, Pechenick et al. 2015, Iliev et al. 2016) or other diachronic corpora (e.g. Hills & Adelman 2015, Snefjella et al. 2018, Kutuzov et al. 2017), as well as other languages (e.g. Viklund & Borin 2016, Hengchen et al. 2019, Marjanen et al. 2020).

While the approach of culturomics relies on investigating change in the usage intensity of words, and especially the data around their first appearances, it should be extended to also consider fluctuations in the meaning that words represent. Tahmasebi & Risse (2017) discuss the utility of automatic sense detection for archive users and digital humanities research. They propose a sense-based approach to capture changes related to the usage and culture of a word. We also believe that correctly recognized shifts in term meanings should be accounted for in order to produce reliable data in any cultural study based on the analysis of the aggregate statistics obtained from term occurrences and term relations over time. Fridlund et al. (2019) attempted to estimate the number of certain types of events (in particular, terror attacks) in the past, portraying the societal responses based on a diachronic document collection (e.g., news archives spanning a longer time period). Inspired by their work, we take as an example a political demonstration. Simply issuing direct queries such as *political demonstration* to a search engine indexing a document archive would be insufficient, and would likely produce inaccurate results. This is because the term *demonstration* and its close derivatives were probably not used in the past to indicate a public show of feelings in support of or against something, or at least one cannot assume this was always the case. Past meanings of *demonstration*, *political demonstration* or *public demonstration* probably did not exactly correspond to their contemporary meanings. Many events that would currently be regarded and labeled as such would be missed during the data collection. In addition, some false negatives can be identified if one does not properly take into consideration diachronic semantic change.[8] On the other hand, equipped with knowledge of actual terms used in the past and accounting for semantic variations in the known term and related ones, the researchers could more accurately collect data and more credibly represent the true frequency of target types of incidents over time. In general, semantic change awareness should improve trustworthy, precise collection building and, by extension, culturomics studies in general.

While the approaches were usually developed for long-term sense tracking and analysis (over decades or centuries), recently researchers have also focused on analyzing diachronic change over shorter time spans such as a few years (Dodds et al. 2011, Danescu-Niculescu-Mizil et al. 2013, Eisenstein et al. 2014, Goel et al. 2016, Del Tredici & Fernández 2018). Short-term change is intensified nowadays due to the popularity of the Web, the high dynamics of social media, and the dramatic increase in the speed of information exchange brought about by novel communication and Web technologies. These all mean that lexical change can

[8] In our example, the sense of *demonstration* as a 'public show of feeling by a number of persons in support of some political or social cause' dates back to 1839 (https://www.etymonline.com/search?q=demonstration).

now materialize in much shorter time frames than in the past. The technologies developed for semantic change analysis over long-span diachronic corpora could be adapted for cultural studies drawing on temporal and social aspects of social media.

2.2 Natural language processing

We expect that many text processing tasks ranging from POS tagging, grammatical dependency detection, semantic role labeling, named entity extraction and linking, and sentiment analysis to language inference could benefit from correct estimation of word senses present in past documents. Currently, post-OCR error detection and correction are among the most common text processing procedures applied to old texts. Automatically detecting and correcting errors in OCR-processed historical texts (Chiron et al. 2017) could also benefit from the research on diachronic conceptual change. This is because knowledge of a word sense that is expected at a given position in a text should help to determine whether the word at that position is erroneous or not (especially in the case of so-called "real-word errors" which are misspellings that result in valid words). This process should also help to generate the most plausible substitutes if the word is deemed an error.

2.3 Document analysis and understanding

We expect that knowledge of change in the diachronic semantics of words constituting a document created at a certain time in the past should help in the analysis of the document (Tahmasebi 2013). Below we discuss three examples.

2.3.1 Providing temporal context to support analysis of past documents

Jatowt et al. (2019) proposed viewing a past document through the lens of its time by utilizing knowledge of the change in the frequency and semantics of words contained in the document (e.g., based on a large diachronic corpus such as Google Books Ngrams). The document in context of its time (DICT) visualization style lets users (e.g., professionals such as historians or other humanities researchers studying old literature) observe whether the words in the document were frequently used or were rather rare at the time of the document's creation. This helps to locate neologisms and archaisms used by the document's author. Furthermore, words that have changed their meaning when compared to a given specified date (e.g., a present time) are identified in text. Together, these functions let users better understand the writing style of the author, which can be important in literature studies, and let them "connect" the document to the words and their senses commonly used at the time the document was created.

2.3.2 Comprehensibility of past texts

Another possible application is to support comprehensibility of past documents. Methods designed to estimate the reading difficulty of past documents could then be incorporated into archival retrieval engines and recommendation systems, so that relevant past texts are provided that current users can understand. Many of the methods described in this book could then be useful, as awareness of changes in word meaning over time could lead to increase the ease of reading and comprehension, as suggested by Tahmasebi & Risse (2013). One application would be to design extensions to traditional readability indexes to cover the additional difficulty caused by the semantic change. Other examples of initiatives in this direction are highlighting words in old texts that have undergone considerable change, as suggested by Jatowt et al. (2019), and/or clarifying their actual senses to improve comprehension by the average reader. The work of Tran et al. (2015) is related to this idea of COMPREHENSION-FOCUSED DOCUMENT ENRICHMENT. They propose recontextualizing past texts by enriching them with explanatory content extracted from Wikipedia, although Wikipedia does not explicitly focus on diachronic conceptual change.

2.3.3 Diachronic text evaluation

Diachronic conceptual change detection and exploration can also be applied to support the date of origin detection. This task is also called DIACHRONIC TEXT EVALUATION (DTE) (Popescu & Strapparava 2015). Many of the DTE solutions rely on information about word occurrence in the past, with the underlying hypothesis that if a document contains many words that were common at a given time point in the past, it is likely that it was created/published at that time point. This is especially the case if the words were rarely used at other time points (see, e.g. Kanhabua & Nørvåg 2009, Chambers 2012, Szymanski & Lynch 2015, Jatowt & Campos 2017). Including information on diachronic conceptual change could further improve the performance of DTE, as demonstrated by Fremann & Lapata (2016) through task-based evaluation of a Bayesian model of diachronic meaning change. This is because additional information on a word's sense (or the probability distribution over its known senses) can be utilized alongside the frequency-based signal to more precisely determine age.

2.4 Information access and recommendation

2.4.1 Query suggestion and content ranking

Semantic change detection techniques can enhance information access and retrieval for users of digital libraries and digital archives. For example, at the query-level, effective query suggestion and correction techniques can be provided based on the computation of across-time analogies and present-to-past semantic relations. These could help users who lack the specific vocabulary of things common in that era to select appropriate query terms. Berberich et al. (2009) and Holzmann et al. (2012) discussed direct application of a method for finding analogical entities across time to information retrieval, mainly for query suggestions.

Recently, some work has aimed at across-time content retrieval and matching to enable novel information access approaches in news archives. For example, semantic term matching was used for extracting and summarizing comparative sentences (Duan & Jatowt 2019), computing temporal analogies (Zhang et al. 2015, Szymanski 2017) or estimating the contemporary relevance of past news articles, defined as the degree of the utility and attractiveness of old news articles to present-day users (Sato et al. 2019). Morsy & Karypis (2016) also propose using information about language change in document similarity computation.

2.4.2 Recommending words with interesting change history

As mentioned earlier, etymological knowledge is not only interesting to professionals such as linguists, historians, or librarians, but also to the wider public. Educators could use it to make students aware of etymological developments and arouse their interest in learning about language and history. Computational approaches and particularly online interactive systems could help to further disseminate knowledge of word etymologies. For such systems to be effective and attractive, it would be beneficial to recommend interesting words to be explored by non-professional users. Explanation of past meanings of words like *gay* or *nice*, for example, tends to surprise lay users who are not aware of them, and triggers questions on the reasons for the change. Existing systems for exploring diachronic conceptual change require users to provide words as the input. As average users may not know what words to search for, recommending sample queries to explore and learn about semantic change could be a useful option to attract or entertain users. Unique or specific input words could be recommended based on the shapes of their self-similarity plots over time (e.g., words that retained stable senses over a long time, or that underwent significant semantic shifts within short time frames) (Jatowt et al. 2018). Finally, suggesting words that show interesting or unexpected semantic changes could be used to provide attractive content on specialized websites, or even might inspire new books directed at the average reader.

2.5 Temporal summarization and trend detection in domain-specific diachronic collections

Detecting shifts in word senses can also be limited to specific domains such as scholarly or legal documents. For example, Degaetano-Ortlieb & Strötgen (2017) analyze differences in frequency, meanings and the underlying temporal scopes of temporal expressions used in scientific writing from 1665 to 2007.

From an application viewpoint, semantic analysis of specialized terminology could help in detecting emerging trends (Dridi et al. 2019) and in summarizing entire domain-specific collections (Mohd Pozi et al. 2020). For example, Dridi et al. (2019) propose detecting emerging trends in scholarly publication collections in computer science and bioinformatics. Rather than employing citation analysis or straightforward frequency-based trend assessment, as has been usually done, the authors use temporal word embeddings to observe shifts in scientific language over several decades. A simple improvement of this approach would be to use contextualized embedding models pre-trained on domain-specific collections such as SciBert, which was trained on scholarly corpora (Beltagy et al. 2019). Another option is to consider specialized term extraction techniques such as ones based on recognizing meaning shifts between general and domain-specific language (Hätty et al. 2019).

A further extension is debiasing semantic change of analyzed words by considering the overall change direction of the collection. Such *temporal normalization in domain-constrained collections* would remove the overall, average drift that the collection underwent over time based on the evolution trajectory of the words studied. This will help to better represent the specific semantic change of these words. Techniques for gender-specific and other kinds of debiasing of word embeddings could be adapted (Bolukbasi et al. 2016, Kaneko & Bollegala 2019).

In general, change detection, temporal summarization, emerging trend detection, and other similar tasks in domain-constrained document collections are promising applications for the computational tools of diachronic semantic change detection and analysis.

3 Conclusions

Recently, we have witnessed many developments and advances in methods for recognizing, analyzing and understanding diachronic semantic change and lexical replacement. In this chapter, we discussed examples and applications of these methods besides the usual purpose of supporting research in historical linguistics by revealing unknown change and improving understanding of known change. We began by surveying representative visual systems that can help the wider public and non-professional users investigate evidences of semantic change and so learn about word etymology and evolution. Finally, we discussed the possibilities of enhancing and improving downstream applications in NLP, information retrieval, and related fields.

Acknowledgements

Thanks go to the Center for Digital Humanities (CDH), University of Gothenburg and to Kyoto University. A first version of this work was written while Nina Tahmasebi was employed at CDH. Adam Jatowt was employed by Kyoto University when the first version was completed. This work has in part been funded by the project *Towards Computational Lexical Semantic Change Detection* supported by the Swedish Research Council (2019–2022; contract 2018-01184), the project *South Asia as a linguistic area?* funded by the Swedish Research Council (2015–2020; contract 2014-00969), and *Nationella språkbanken* (the Swedish National Language Bank), a research infrastructure jointly financed by the Swedish Research Council (2018–2024; contract 2017-00626) and the 10 partner institutions. In addition we are grateful for the strong and enduring commitment that the University of Gothenburg, its Faculty of Humanities, and its Department of Swedish have shown to our R&D unit Språkbanken Text for almost 50 years. The authors would also like to acknowledge Dominik Schlechtweg, Haim Dubossarsky, and Simon Hengchen for fruitful discussions and insights. The authors wish to thank the Department of Literature, History of Ideas, and Religion at the University of Gothenburg for providing financial support for language editing services.

Abbreviations

ATAR	aspect-based temporal analog retrieval
IR	information retrieval
KWIC	keyword in context
LSA	latent semantic analysis
NLP	natural language processing
OCR	optical character recognition
PCA	principal component analysis
POS	part of speech
YAGO	Yet Another Great Anthology

References

Acerbi, Alberto, Vasileios Lampos, Philip Garnett & R. Alexander Bentley. 2013. The expression of emotion in 20th century books. *PLOS ONE* 8(3). e59030.

Beltagy, Iz, Kyle Lo & Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Bentley, R. Alexander, Alberto Acerbi, Paul Ormerod & Vasileios Lampos. 2014. Books average previous decade of economic misery. *PLOS ONE* 9(1). e83147.

Berberich, Klaus, Srikanta J. Bedathur, Mauro Sozio & Gerhard Weikum. 2009. *Bridging the terminology gap in Web archive search*. Paper presented at the Twelfth International Workshop on the Web and Databases (WebDB 2009). http://webdb09.cse.buffalo.edu/papers/Paper20/webdb2009-final.pdf.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama & Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Bradley, Margaret M. & Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25(1). 49–59.

Carpendale, Sheelagh. 2008. Evaluating information visualizations. In Andreas Kerren, John T. Stasko, Jean-Daniel Fekete & Chris North (eds.), *Information Visualization: Human-centered issues and perspectives*, 19–45. Berlin: Springer.

Chambers, Nathanael. 2012. Labeling documents with timestamps: Learning from their time expressions. In *ACL 2012*, 98–106.

Chiron, Guillaume, Antoine Doucet, Mickaël Coustaty & Jean-Philippe Moreux. 2017. ICDAR2017 competition on post-OCR text correction. In *Document analysis and recognition (ICDAR), 2017*, vol. 1, 1423–1428.

Collins, Christopher, Gerald Penn & Sheelagh Carpendale. 2008. Interactive visualization for computational linguistics. In *Proceedings of HLT 2008: Tutorial abstracts*, 6–6. ACL.

Cox, Michael A.A. & Trevor F. Cox. 2008. Multidimensional scaling. In Chunhouh Chen, Wolfgang Karl Härdle & Antony Unwin (eds.), *Handbook of data visualization*, 315–347. Heidelberg: Springer.

Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec & Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of WWW 2013*, 307–318. ACM.

Degaetano-Ortlieb, Stefania & Jannik Strötgen. 2017. Diachronic variation of temporal expressions in scientific writing through the lens of relative entropy. In *International Conference of the German Society for Computational Linguistics and Language Technology*, 259–275. Springer.

Del Tredici, Marco & Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of COLING 2018*, 1591–1603. ACL.

Dodds, Peter Sheridan, Kemeron Decker Harris, Catherine A. Bliss & Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLOS ONE* 6(12). e26752.

Dridi, Amna, Mohamed Medhat Gaber, R. Muhammad Atif Azad & Jagdev Bhogal. 2019. Leap2trend: A temporal word embedding approach for instant detection of emerging scientific trends. *IEEE Access* 7. 176414–176428.

Duan, Yijun & Adam Jatowt. 2019. Across-time comparative summarization of news articles. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 735–743.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith & Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE* 9(11). e113114.

Frermann, Lea & Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the ACL* 4. 31–45. DOI: 10.1162/tacl_a_00081.

Fridlund, Mats, Leif-Jöran Olsson, Daniel Brodén & Lars Borin. 2019. Trawling for terrorists: A big data analysis of conceptual meanings and contexts in Swedish newspapers. In *5th International Workshop on Computational History, HistoInformatics@TPDL 2019* (CEUR Workshop Proceedings 2461), 30–39. http://ceur-ws.org/Vol-2461/paper_5.pdf.

Goel, Rahul, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz & Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, 41–57. Springer.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of EMNLP 2016*, 2116–2121. Austin: ACL. DOI: 10.18653/v1/D16-1229.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Hätty, Anna, Dominik Schlechtweg & Sabine Schulte im Walde. 2019. SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, 1–8.

Hellrich, Johannes, Sven Buechel & Udo Hahn. 2018. JeSeme: A website for exploring diachronic changes in word meaning and emotion. *CoRR* abs/1807.04148. http://arxiv.org/abs/1807.04148.

Hengchen, Simon, Ruben Ros & Jani Marjanen. 2019. A data-driven approach to the changing vocabulary of the *nation* in English, Dutch, Swedish and Finnish newspapers, 1750–1950. In *Proceedings of the Digital Humanities (DH) conference 2019*.

Heylen, Kris, Dirk Speelman & Dirk Geeraerts. 2012. Looking at word meaning: An interactive visualization of semantic vector spaces for Dutch synsets. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 16–24. ACL.

Hills, Thomas T. & James S. Adelman. 2015. Recent evolution of learnability in American English from 1800 to 2000. *Cognition* 143. 87–92.

Hilpert, Martin & Stefan Th. Gries. 2008. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24(4). 385–401.

Hilpert, Martin & Florent Perek. 2015. Meaning change in a petri dish: Constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard* 1(1). 339–350.

Holzmann, Helge, Gerhard Gossen & Nina Tahmasebi. 2012. Fokas: Formerly known as, a search engine incorporating named entity evolution. In *Proceedings of COLING 2012: Demonstration papers*, 215–222.

Iliev, Rumen, Joe Hoover, Morteza Dehghani & Robert Axelrod. 2016. Linguistic positivity in historical texts reflects dynamic environmental and psychological factors. *PNAS* 114(2). E7871–E7879.

Jatowt, Adam & Marc Bron. 2016. HistoryComparator: Interactive across-time comparison in document archives. In *Proceedings of COLING 2016*, 84–88.

Jatowt, Adam & Ricardo Campos. 2017. Interactive system for reasoning about document age. In *Proceedings of CIKM 2017*, 2471–2474. ACM.

Jatowt, Adam, Ricardo Campos, Sourav S. Bhowmick & Antoine Doucet. 2019. Document in context of its time (DICT): Providing temporal context to support analysis of past documents. In *Proceedings of CIKM 2019*, 2869–2872. Beijing: ACM.

Jatowt, Adam, Ricardo Campos, Sourav S. Bhowmick, Nina Tahmasebi & Antoine Doucet. 2018. Every word has its history: Interactive exploration and Visualization of word sense evolution. In *Proceedings of CIKM 2018*, 1899–1902. ACM.

Jatowt, Adam & Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of Joint Conference on Digital Libraries* (JCDL '14), 229–238. http://dl.acm.org/citation.cfm?id=2740769.2740809.

Kalouli, Aikaterini-Lida, Rebecca Kehlbeck, Rita Sevastjanova, Katharina Kaiser, Georg A. Kaiser & Miriam Butt. 2019. ParHistvis: Visualization of parallel multilingual historical data. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 109–114.

Kaneko, Masahiro & Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv preprint arXiv:1906.00742.*

Kanhabua, Nattiya & Kjetil Nørvåg. 2009. Using temporal language models for document dating. In *Machine learning and knowledge discovery in databases*, 738–741. Springer.

Kenter, Tom, Melvin Wevers, Pim Huijnen & Maarten De Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1191–1200.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore: ACL. DOI: 10.3115/v1/W14-2517.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey, Erik Velldal & Lilja Øvrelid. 2017. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the events and stories in the news workshop*, 31–36. ACL.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449(7163). 713.

Marjanen, Jani, Jussi Kurunmäki, Lidia Pivovarova & Elaine Zosa. 2020. The expansion of isms, 1820–1917: Data-driven analysis of political language in digitized newspaper collections. *Journal of Data Mining & Digital Humanities.* https://jdmdh.episciences.org/6728.

Martinez-Ortiz, Carlos, Tom Kenter, Melvin Wevers, Pim Huijnen, Jaap Verheul & Joris Van Eijnatten. 2016. Design and implementation of ShiCo: Visualising shifting concepts over time. In *HistoInformatics 2016*, vol. 1632, 11–19.

Mazeika, Arturas, Tomasz Tylenda & Gerhard Weikum. 2011. Entity timelines: Visual analytics and named entity evolution. In *Proceedings of the 20th ACM international conference on information and knowledge management*, 2585–2588. ACM.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331. 176–182.

Mohd Pozi, Muhammad Syafiq, Adam Jatowt & Yukiko Kawai. 2020. Temporal summarization of scholarly paper collections by semantic change estimation: Case study of CORD-19 dataset. In *JCDL 2020*, 459–460. Virtual Event, China: ACM.

Morsy, Sara & George Karypis. 2016. Accounting for language changes over time in document similarity search. *ACM Transactions on Information Systems* 35(1).

Odijk, Daan, Giuseppe Santucci, Maarten de Rijke, Marco Angelini, Guido Lorenzo Granato, et al. 2014. Time-aware exploratory search: Exploring word meaning through time. In *Sigir 2012 workshop on time-aware information access.*

Pagel, Mark, Quentin D. Atkinson & Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163). 717.

Pechenick, Eitan Adam, Christopher M. Danforth & Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE* 10(10). e0137041.

Popescu, Octavian & Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *SemEval 2015*, 870–878.

Rohrdantz, Christian, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim & Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of HLT 2011: Short papers*, 305–310. ACL.

Sato, Mari, Adam Jatowt & Masatoshi Yoshikawa. 2019. Present relatedness estimation of archival documents using external knowledge base. In *DEIM forum 2019*, 1–5. https://db-event.jpn.org/deim2019/post/papers/275.pdf.

Schätzle, Christin & Miriam Butt. 2020. Visual analytics for historical linguistics: Opportunities and challenges. working paper or preprint. https://hal.inria.fr/hal-02914284.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of SemEval 2020*, 1–23. Barcelona: ACL. https://www.aclweb.org/anthology/2020.semeval-1.1.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–123.

Snefjella, Bryor, Michel Généreaux & Victor Kuperman. 2018. Historical evolution of concrete and abstract language revisited. *Behavior Research Methods* online first. 1–13. DOI: 10.3758/s13428-018-1071-2.

Szymanski, Terrence. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of ACL 2017*. ACL.

Szymanski, Terrence & Gerard Lynch. 2015. UCD: Diachronic text classification with character, word, and syntactic n-grams. In *SemEval 2015*, 879–883.

Tahmasebi, Nina. 2013. *Models and algorithms for automatic detection of language evolution*. Gottfried Wilhelm Leibniz Universität Hannover. (Doctoral dissertation). http://edok01.tib.uni-hannover.de/edoks/e01dh13/771705034.pdf.

Tahmasebi, Nina & Thomas Risse. 2013. The role of language evolution in digital archives. In *SDA 2013*, vol. 1091, 16–27. http://ceur-ws.org/Vol-1091/paper2.pdf.

Tahmasebi, Nina & Thomas Risse. 2017. On the uses of word sense change for research in the digital humanities. In *TPDL 2017*, 246–257. Berlin: Springer.

Tang, Xuri. 2018. Survey paper: A state-of-the-art of semantic change computation. *Natural Language Engineering* 24(5). 649–676.

Theron, Roberto & Laura Fontanillo. 2015. Diachronic-information visualization in historical dictionaries. *Information Visualization* 14(2). 111–136.

Tran, Nam Khanh, Andrea Ceroni, Nattiya Kanhabua & Claudia Niederée. 2015. Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization. In *Proceedings of WSDM 2015*, 339–348. Shanghai: ACM. DOI: 10.1145/2684822.2685315.

van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9. 2579–2605.

van Eijnatten, Joris, Toine Pieters & Jaap Verheul. 2014. Using texcavator to map public discourse. *Tijdschrift voor Tijdschriftstudies* 35. 59–65.

Viklund, Jon & Lars Borin. 2016. How can big data help us study rhetorical history? In *Selected Papers from the CLARIN Annual Conference 2015*, 79–93. Linköping: LiU EP. http://www.ep.liu.se/ecp/123/007/ecp15123007.pdf.

Xu, Zaikun & Fabio Crestani. 2017. Temporal semantic analysis and visualization of words. In *IIR*, 52–62.

Zhang, Yating, Adam Jatowt, S. Sourav Bhowmick & Yuji Matsumoto. 2019. ATAR: Aspect-based temporal analog retrieval system for document archives. In *WSDM2019*. Melbourne: ACM.

Zhang, Yating, Adam Jatowt, Sourav Bhowmick & Katsumi Tanaka. 2015. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proceedings of ACL/IJCNLP 2015 (Volume 1: Long papers)*, 645–655. Beijing: ACL. DOI: 10.3115/v1/P15-1063.

Zhang, Ying, Stephan Vogel & Alex Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of LREC 2004*. ELRA.

Chapter 11

Challenges for computational lexical semantic change

Simon Hengchen[a], Nina Tahmasebi[a], Dominik Schlechtweg[b] & Haim Dubossarsky[c]

[a]University of Gothenburg [b]University of Stuttgart [c]University of Cambridge

The computational study of lexical semantic change (LSC) has taken off in the past few years and we are seeing increasing interest in the field, from both computational sciences and linguistics. Most of the research so far has focused on methods for modelling and detecting semantic change using large diachronic textual data, with the majority of the approaches employing neural embeddings. While methods that offer easy modelling of diachronic text are one of the main reasons for the spiking interest in LSC, neural models leave many aspects of the problem unsolved. The field has several open and complex challenges. In this chapter, we aim to describe the most important of these challenges and outline future directions.

1 Introduction

The goal of tackling lexical semantic change (LSC) computationally is primarily to reconstruct semantic change evident in large diachronic corpora. The first papers addressing LSC appeared in 2008–2009 and since then a few papers per year have been published.[1] The first works that used neural embeddings were published in 2014 and 2015 (Kim et al. 2014, Kulkarni et al. 2015, Dubossarsky et al. 2015) and together with Hamilton et al. (2016b), they sparked interest in the

[1] “Language evolution”, “terminology evolution”, “semantic change”, “semantic shift” and ’‘semantic drift” are all terms that are or have been used for the concept which we denote *lexical semantic change.*

Simon Hengchen, Nina Tahmasebi, Dominik Schlechtweg & Haim Dubossarsky. 2021. Challenges for computational lexical semantic change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu & Simon Hengchen (eds.), *Computational approaches to semantic change*, 341–372. Berlin: Language Science Press. DOI: 10.5281/zenodo.5040322

research community in the problem of LSC.[2] Although there are a few research groups that have a longer history of studying LSC using computational methods, the majority of papers are single-entry papers where a group with an interesting model apply their method to a novel application on popular diachronic data. This leads to quick enhancement of methods but limits development and progress in other aspects of the field.

When surveying prior work, it is obvious that the computational field of LSC has been divided into two strands. The first strand deals with words as a whole and determines change on the basis of a word's dominant sense (e.g. Kim et al. 2014, Kulkarni et al. 2015). An oft-used example is *gay*[3] shifting from its 'cheerful' sense to 'homosexual'. The second strand deals with a word's senses[4] individually – for example, the 'music' sense of *rock* has gradually come to describe not only music but also a certain lifestyle, while the 'stone' sense remained unchanged (as seen in the works of Tahmasebi 2013 and Mitra et al. 2015). The first strand took off with the introduction of neural embeddings and its easy modelling of a word's semantic information. The second strand, faced with the immense complexity of explicitly modelling senses and meaning, has received much less attention.

Computational models of meaning are at the core of LSC research, regardless of which strand is chosen. All current models, with the exception of those purely based on frequency, rely on the distributional hypothesis, which brings with it the set of challenges discussed in Section 3. But even accepting the distributional hypothesis and assuming meaning in context, the problem formulated by Schütze (1998) remains: how does one accurately portray a word's senses? The question is valid regardless of whether the senses are represented individually or bundled up into one single representation. Recent developments in contextual embeddings (e.g. Peters et al. 2018) provide hope for accurate modelling of senses. However, they do not alleviate the problem of grouping sentence representations into sense correspondence. Within natural language processing (NLP), computational models of word meaning are often taken at face value and not questioned

[2] Compare, for example, the roughly 30 papers at the start of 2018 as reported by Tahmasebi et al. (2018), with the roughly 50 papers submitted at the 1st International workshop on computational approaches to historical language change 2019 (Tahmasebi et al. 2019), and recent submissions at xACL venues, including the 21 papers submitted to the SemEval-2020 Task 1 on unsupervised lexical semantic change detection (Schlechtweg et al. 2020).

[3] More often than not, parts of speech are collapsed – in this case, there is thus no difference between the adjective and the noun.

[4] For the sake of clarity we use this as a simplified wording. We do not imply that a fixed number of senses exist in a sense inventory; instead senses can overlap and be assigned different strengths.

by researchers working on LSC. This is thus one of the areas that needs further attention in future work. Another area for thorough investigation is how useful sense-differentiation is for accurate LSC models.

Another important area for future work is robust evaluation. Computational LSC methods model textual data as information signals and detect change in these signals. A signal can be a multidimensional vector, a cluster of words, topics, or frequency counts. This increasing level of abstraction is often ignored in evaluation; current evaluation standards allow for anecdotal evaluation of signal change, often without tying the results back to the text. Can we find evidence in the text for the detected changes?[5] So far, semantic annotation is the only way to evaluate methods on historical corpora while making sure that expected changes are present in the text. Annotating involves a significant investment of time and funds, and results in a limited test set. A middle ground is to evaluate with respect to an outside source, like a dictionary or encyclopedia. However, while these resources offer an "expected" time and type of change, we can never be certain that these changes are reflected in the corpus under study. We refer to the example of *computer* in Tahmasebi et al. (2018): the different parts of Google Books (British English, American English, and German) that reach the same level of frequency in different periods in time, 1934 for the German portion, 1943 for the American English, and 1953 for the British English. Recent work (Kulkarni et al. 2015, Dubossarsky et al. 2019, Shoemark et al. 2019, Schlechtweg & Schulte im Walde 2020) introduced relatively cheap methods of generating *synthetic* semantic change for any dataset, which we believe is an important path forward. The question is not if, but how synthetic evaluation data can complement costly manual evaluation. This will be answered in Section 4.

In addition, current methods work with rather coarse time granularity, largely because of the inherent complexity of adding multiple time bins (and senses) to the models. Unfortunately this constraint limits both the possibilities of the methods, and the results that can be found. Again, adding complexity to the models results in complexity in evaluation and calls for robust evaluation methods and data.

In this chapter, we will discuss current and future challenges, and outline avenues for future work. More specifically, Section 2 discusses the requirements for textual resources, and their role for LSC. Section 3 covers models of meaning,

[5] To the best of our knowledge, only Hengchen (2017) evaluates semantic change candidates output by a system reading a relatively large sample of sentences from the corpus studied – but only for a single word, while several projects make use of extensive annotation to ensure that detected changes are present in the underlying textual corpus (Lau et al. 2012, Schlechtweg et al. 2017, 2018, 2020, Hätty et al. 2019, Perrone et al. 2019, Giulianelli et al. 2020).

the current limitations of computational models, models of change, and what remains to be done in terms of time and sense complexity. Section 4 sheds light on the need for robust evaluation practices and what we have learned so far. We then proceed in Section 5 to showing the potential for collaboration between computational LSC and other fields, and end with some concluding remarks.

2 Data for detecting LSC

Hand in hand with the fast and simple modelling of word meaning, using neural embeddings for example, is the easy access to digital, diachronic texts that sparked mainstream interest in LSC as a problem domain for testing new models. For many reasons, including the early availability of large English corpora, there has long been a large over-representation of studies performed on English, in particular using COHA (Davies 2002), Google N-grams (Michel et al. 2011), and various Twitter corpora (see Table 2 in Tahmasebi et al. 2018 for an overview). As a consequence, most of the computational modelling of LSC has been developed and evaluated on these resources. However, tools and methods developed on one language (e.g., English) are not easily transferable to another language, a reoccurring challenge in other fields of NLP as well (Bender 2011, Ponti et al. 2019). Moreover, many languages may even lack the amount of historical, digitized data needed for robustly employing state-of-the-art methods like neural embeddings. This point is reinforced if we follow the recommendations of Bowern (2019) for example, and distinguish groups based on features such as location, age, and social standing. The immediate result of this limitation is that many languages remain unstudied, or worse, studied with unsuitable methods. Consequently, whatever conclusions are drawn about LSC and formulated as “laws” are based on a very limited sample of languages, which may not be representative of the other 7,100 living languages.[6] As LSC mainly focuses on diachronic text, one obvious area for research is determining how well methods that rely on representations developed primarily for modern text transfer to historical languages.[7]

[6] Figure from ethnologue.com, https://www.ethnologue.com/guides/how-many-languages, last accessed 2020/01/16, rounded down to the lower hundred. In addition to living languages, dead languages are also studied, e.g. Rodda et al. (2017), Perrone et al. (2019), or McGillivray et al. (2019) for Ancient Greek.

[7] See Piotrowski (2012) for a thorough overview of such challenges, and Tahmasebi et al. (2013) for an example of the applicability of a cluster-based representation on historical data.

An important criterion for textual resources for LSC is good, reliable time stamps for each text. The texts should also be well distributed over longer time periods.[8] For these reasons, newspaper corpora are popular for LSC studies. They also have the advantage that different news items are roughly equal in length. But while they cover a large and interesting part of society, including technological inventions, they are limited in their coverage of everyday, non-newsworthy life.[9] On the other hand, large literary corpora like the Google N-grams pose different challenges including their skewed sampling of topics over time. For example, Pechenick et al. (2015) point to an over-representation of scientific literature in the Google Books corpus, which biases the language used toward specific features mostly present in academic writing. A second challenge to the use of the Google N-grams corpus is the small contexts (at most five words in a row) and the scrambled order in which these contexts are presented. To what extent this large resource can be used to study LSC remains to be investigated. One important question is whether changes found can be thoroughly evaluated using only these limited contexts.[10]

Other, less known corpora have other known deficits. For example, many literary works come in multiple editions with (minor) updates that modernize the language, while other texts lack known timestamps or publication dates. Some authors are more popular than others (or were simply more productive) and thus contribute in larger proportion and risk skewing the results.[11]

What an optimal resource looks like is clearly dependent on the goal of the research, and LSC research is not homogeneous; different projects have different aims. While some aim to describe an interesting dataset (like the progression of one author), others want to use large-scale data to generalize to language outside of the corpus itself. In the latter case, it is important to be varied, as large textual

[8] Recurring advertisements running for weeks at a time can effectively bias the corpus. See for example, the work by Prescott (2018: 67) for a case study on the Burney newspaper collection.

[9] Similar to news corpora, Twitter, another popular source for LSC research, offers posts which have timestamps and are consistent in size (though radically shorter than news articles), but with very different characteristics. However, most Twitter corpora are short-term, unlike the longer temporal dimensions of many news corpora.

[10] Although there are several LSC papers using Google N-grams, e.g., Wijaya & Yeniterzi (2011) or Gulordava & Baroni (2011), to date there are no systematic investigations into the possibility of detecting different kinds of LSC, nor any systematic evaluation using grounding of found change in the Google N-grams.

[11] See, for example, Tangherlini & Leonard (2013) where a single book completely changed the interpretation of a topic. Similarly, one can find many editions and reprints of the Bible in the Eighteenth Century Collections Online (ECCO) dataset that spans a century and contains over 180,000 titles; which will influence models. For a study on the effects of text duplication on semantic models, see Schofield, Thompson, et al. (2017).

corpora are not random samples of language as a whole (see, e.g., Koplenig 2016) and whatever is encountered in corpora is only valid for those corpora and not for language in general.

Most existing diachronic corpora typically grow in volume over time. Sometimes this stems from the amount of data available. At other times it is an artefact related to ease of digitisation (e.g., only certain books can be fed into an automatic scanner) and OCR technology (OCR engines are trained on specific font families). This growth results in an extension of vocabulary size over time, which might not reflect reality and has serious effects on our methods. For example, previous work has shown that diachronic embeddings are very noisy (Hellrich & Hahn 2016, Dubossarsky et al. 2017, 2019, Kaiser et al. 2020, Schlechtweg et al. 2020) with a large frequency bias and are clearly affected by more and more data over time. Current and future studies are thus left with the question of whether the signal change we find really correspond to LSC in the text, or whether it is simply an artefact of the corpus.

We can only find what is available in our data: if we want to model other aspects of language and LSC, we need datasets that reflect those aspects well (for similar considerations related to evaluation, see Section 4.1.1). This fact makes the case for using texts stemming from different sources, times, and places to allow for (re-)creating the complex pictures of semantic change. Thus general aspects beneficial for LSC are texts that are well-balanced (in time and across sources) and high-quality (with respect to OCR quality) with clear and fine-grained temporal metadata, as well as other kinds of metadata that can be of use. Until now, most existing computational LSC studies have been performed on textual data exclusively, aside from Perrone et al. (2019) and Jawahar & Seddah (2019) who respectively used literary genre and social features as features. The reason for the under-utilisation of extra-linguistic metadata – despite there being great need for it, as advocated by Bowern (2019) – is to a large extent the lack of proper and reliable metadata. In the case of Google N-grams, this kind of metadata is sacrificed in favour of releasing large volumes of data freely. This path is also promising with respect to modelling the individual intent, described in Section 3.3.1.

For the long-term future, we should raise the question of whether we can model language at all using only texts (Bender & Koller 2020). How much can we improve with multi-modal data in the future (Bruni et al. 2012), and what kind of data would be beneficial for LSC?

3 Models of meaning and meaning change

In this section, we shed light on the "meaning" we strive to model from the data, and on the challenges involved with modelling meaning computationally, and finally on how to employ the resulting information signals to establish change.

3.1 Theory of lexical meaning and meaning change

The field urgently needs definitions of the basic concepts it wants to distinguish: after all, we can draw from a rich tradition of semantics and semantic change research. The field traditionally starts with Reisig (1839), although Aristotle (1898) theorized metaphors in his *Poetics* well before then.[12]

Here we focus on one theory which encompasses many others. Blank (1997: 54) distinguishes three different levels of word meaning based on which type of knowledge a word can trigger in a human: (i) language-specific semantic, (ii) language-specific lexical, and (iii) language-external knowledge. The first comprises core semantic knowledge needed to distinguish different word meanings from each other.[13] This knowledge corresponds to the minimal language-specific semantic attributes needed to structure a particular language, often called "sememe" in structural semantics. From these follow the hierarchical lexical relations between words (e.g. synonymy or hypernymy). The second level of word meaning comprises knowledge about the word's role in the lexicon (part of speech, word family or knowledge about polysemy/multiple meanings, referred to as "senses" in this chapter). It includes the rules of its use (regional, social, stylistic or diachronic variety; syntagmatic knowledge such as selectional restrictions, phraseologisms or collocations). Level (iii) comprises knowledge about connotation and general knowledge of the world.

Blank (1997) assumes that the knowledge from these three levels is stored in the mental lexicon of speakers, which can also change historically in these three levels (at least). An example of a change at the language-specific semantic level (i) is Latin *pipio* 'young bird' > 'young pigeon' which gained the attribute [pigeon-like] (Blank 1997: 106–107). An example of change on the language-specific lexical level (ii) is *gota* 'cheek' which changes from being commonly used in Old Italian to being used exclusively in the literary-poetic register in New Italian (Blank 1997: 107). Finally, a change at the language-external knowledge level (iii) occurs when the knowledge about the referent changes. This can occur, for example, when the

[12] See for example the work by Magué (2005) for an overview.

[13] Note that this level covers only what Blank (1997) calls "knowledge" (p. 94). He then distinguishes six further levels of "meaning" (pp. 94–96).

referent itself changes such as with German *Schiff* 'ship', as ships were primarily steamships in the 19th century, while today they are mainly motor ships (Blank 1997: 111).

Unfortunately, as will be made clear in the following subsection, this rich tradition of work is not used by the computational side of LSC because it is difficult to model meaning purely from written text. Currently, our modelling is very blunt. It can primarily capture contextual similarity between lexical items, and rarely distinguishes between different levels of meaning. Whether we draw from the large existing body of work that exists in traditional semantics research, or start from scratch with a new definition of what computational meaning is, we hope researchers in our field can come together and agree on what is, and should, be modelled.

Similar to the conundrum in the definition of word meaning above, studies on LSC detection are seldom clear on the question of which type of information they aim to detect change in (Schlechtweg & Schulte im Walde 2020). There are various possible applications of LSC detection methods (e.g. Hamilton et al. 2016a, Voigt et al. 2017, Kutuzov et al. 2017, Hengchen et al. 2019). Change at different levels of meaning may be important for different applications. For example, for literary studies it may be more relevant to detect changes of style, for social sciences the relevant level may be language-external knowledge and for historical linguistics the language-specific lexical and semantic levels may be more important. Furthermore, LSC can be further divided into types (e.g., broadening/narrowing, amelioration/pejoration, metaphor and hyperbole). Several taxonomies of change have been suggested over the decades (Bréal 1897, Bloomfield 1933 and Blank 1999, to name a few). Clearly, none of these applications or types of change can be properly tested until an adequate model of meaning is developed and the types of LSC to be investigated are meticulously defined.

3.2 Computational models of meaning

The need to choose the textual data available for the models and the decisions regarding the preprocessing of the text are common to all models of computational meaning. While the influence of the former was described in Section 2, and is fairly straightforward, extremely little attention is paid to preprocessing although its effects on the end results are far-reaching. The lower-casing of words often conflates parts of speech. For example *Apple* (proper noun) and *apple* (common noun) cannot be distinguished after lower-casing. Filtering out different parts of speech is also common practice, and can have radical effects on the results.[14] Thus the effects of preprocessing on meaning representations should be investigated in the future.

Nevertheless, the core of studying LSC computationally is the choice of the computational model of meaning: what we can model determines what change we can find. Crucially, methods for computational meaning inherit the theoretical limitations discussed in Section 3.1. The challenge becomes even more cumbersome as existing methods for computational meaning rely on the distributional hypothesis (Harris 1954), which represents word meaning based on the context in which words appear. In so doing, they often conflate lexical meaning with cultural and topical information available in the corpus used as a basis for the model. These limitations are not specific to semantic change, and lie at the basis of a heated debate that questions the fundamental capacity of computational models to capture meaning using only textual data, see for example Bender & Koller (2020).

There are different categories of computational models for meaning. These comprise a hierarchy with respect to the granularity of their sense/topic/concept representations:

(a) a single representation for a word and all its semantic information (e.g., static embeddings),

(b) a representation that splits a word into semantic areas (roughly) approximating senses (e.g., topic models), and

(c) a representation that models every occurrence of a word individually (e.g., contextual embeddings) and possibly groups them post-hoc into clusters of semantically-related uses expressing the same sense.

These categories of models differ with respect to their potential to address various LSC problems. For example, novel senses are hard to detect with models of category (a). However, these models have the upside of producing representations for all the words in the vocabulary, which is not the case for all models in category (b) (see for example Tahmasebi et al. 2013). In contrast, some sense-differentiated methods (category (b)), such as topic modelling allow for easy disambiguation so that we can deduce which word was used in which sense. However, category (a) models (e.g., word2vec) do not offer the same capability as

[14] For discussions on the effects of preprocessing (or, as coined by Thompson & Mimno 2018, "purposeful data modification") for text mining purposes, we refer to Schofield & Mimno (2016), Schofield, Magnusson, et al. (2017), Denny & Spirling (2018), and Tahmasebi & Hengchen (2019).

they provide one vector per word, which is also biased toward the word's more frequent sense.[15]

Furthermore, models that derive meaning representations that can be interpreted and understood are needed to determine which senses of a word are represented, and whether they capture standard word meaning, topical use, pragmatics, or connotation (i.e., to distinguish between the levels of meaning referred to in Section 3.1). The interpretability also allows us to qualitatively investigate different representations to determine which is better for different goals.

Finally, the data requirements of our models can pose a critical limitation on our ability to model meaning (see Section 2) as the computational models of meaning are data hungry and require extremely large amounts of text. They cannot be applied to the majority of the world's existing written languages as those often do not have sufficient amounts of written historical texts. If we follow the proposal of Bowern (2019) and divide our data, not only by time, but also according to social aspects (to e.g. echo Meillet 1905), we reduce the amount of available data even further.

3.3 Computational models of meaning change

Change is defined and computed by comparing word representations between two or more time points, regardless of the specific model of meaning. Different models entail different mathematical functions to quantify the change. For example, and without claiming to be exhaustive: cosine or Euclidean distances are used for embedding models of continuous vectors representations, Hellinger distance and Jensen-Shannon or Kullback-Leibler divergences for topic distributions, and Jensen-Shannon divergence or cross-entropy for sense-differentiated representations. Ultimately the mathematical functions provide only a scalar that represents the degree of change. Determining change type from this information is not straightforward. This impedes our ability to derive fine-grained information about the nature of change, as touched upon in Section 3.1, and to incorporate theories of change which, for instance, postulate direction of change. For example, it becomes difficult to detect which sense changed, or to provide relevant distinctions related to the different applications (e.g., change in meaning vs. change in connotation), or taxonomies of change (e.g., broadening vs. narrowing). Schlechtweg & Schulte im Walde (2020) identified two basic notions of change that are

[15] Every use of the word in a sentence is not accurately described by the vector representing it. In modern texts not pertaining to geology, a vector representation of *rock* is biased toward its more frequent 'music' sense and will be a worse representation of a sentence where *rock* is used in a 'stone' sense.

used in LSC research to evaluate the models' scalar change scores: (i) GRADED LSC, where systems output to what degree words change (Hamilton et al. 2016a, Dubossarsky et al. 2017, Bamler & Mandt 2017, Rosenfeld & Erk 2018, Rudolph & Blei 2018, Schlechtweg et al. 2018), and (ii) BINARY LSC, where systems make a decision on whether words have changed or not (Cook et al. 2014, Tahmasebi & Risse 2017, Perrone et al. 2019, Shoemark et al. 2019). Despite this limitation of coarse scores of LSC, several change types have been targeted in previous research (Tahmasebi et al. 2018: Table 4). Importantly, in order for the results of LSC methods to be valuable for downstream tasks, we see a great need to determine the kind of change (e.g., broadening, narrowing, or novel sense). Methods that only detect one class, namely *changed*, defer the problem to follow-up tasks: in which way has a word changed, or on what level (i)–(iii)[16] from Section 3.1 did the change occur?

3.3.1 Discriminating individual sentences

Meaning is ascribed to words at the sentence (utterance) level. However, for technical reasons related to the limitations of current computational models, previous work has carried out LSC only in large corpora. As a result, we model each word of interest with a signal (topic, cluster, vector) across all sentences and detect change in the signal. This discrepancy between the level at which the linguistic phenomena occur and the level of the analysis that is carried out may account for the type of questions commonly asked in contemporary research. In the majority of the cases, the signal change is evaluated on its own, and the question *did the word meaning change or not?* is the only one answered. In a few rare cases, change is tied to the text and verified using the text. *Did the word change in the underlying corpus or not?* is in fact a much more accurate question but is asked much less frequently. In a future scenario, where our models of computational meaning are much more fine-grained, we will be able to ask a third question: *Is a specific usage of a word different than its previous uses?* To be able to tie the detected changes back to individual usage is much more demanding of any system and requires word sense disambiguation (WSD) to be fully solved. Although radically more challenging, this task is also much more rewarding. It can help us in proper search scenarios, in dialogue and interaction studies, argument mining (where a person's understanding of a concept changes during the conversation), and in literary studies, to name but a few examples.

[16] Though level (iii) relates to change in world-knowledge and goes well beyond semantic change.

3.3.2 Modelling of time

The modelling of meaning change is directly dependent on the time dimension inherent in the data. Often, we artificially pool texts from adjacent years into long time bins because our computational models require large samples of text to produce accurate meaning representations or, to draw from research in historical sociolinguistics, because bins of a certain length are considered as "generations" of language users (Säily 2016). Unfortunately, this leads to loss of fine-grained temporal information. From a modelling perspective, the inclusion of such information has the clear advantage of leading to more ecological models for LSC. This advantage can be used in two main ways: either to mitigate the noise associated with meaning representation models, or to detect regular patterns of change. Understandably, these advantages are only available when sufficient time points are included in the analysis. More time points, however, undoubtedly lead to greater computational complexity – linearly if we consider the comparison of only subsequent time points, or quadratically if we consider all pairwise comparisons.

Some theories of LSC assume that change unfolds gradually through time, creating a trajectory of change (e.g., the regular patterns of semantic change in Traugott & Dasher 2001). Only models that acquire a meaning representation at several time points (e.g. Tsakalidis & Liakata 2020) are able to validate this underlying assumption by demonstrating a gradual trajectory of change. The work by Rosenfeld & Erk (2018) is an interesting example, as it models semantic change as a continuous variable and can also output the rate of change. Good extensions include allowing different change rates for different categories of words, or including background information about time periods where things change differently. In addition, models with multiple time points may contribute to improved LSC modelling by facilitating the discovery of intricate change patterns that would otherwise go unnoticed. For example, Shoemark et al. (2019) analysed Twitter data with high temporal resolution, and reported that several words demonstrated repeating seasonal patterns of change. The analysis of LSC trajectories easily lends itself to the use of modern change detection methods, which holds great promise for detecting hidden patterns of both change and regularities.

4 Evaluation

Thus far, evaluation of LSC methods has predominantly ranged from a few anecdotally discussed examples to semi-large evaluation on (synthetic or pre-compiled) test sets, as made clear by Table 2 in Tahmasebi et al. (2018).[17] The SemEval-2020 Task 1 on unsupervised lexical semantic change detection provided the first larger-scale, openly available dataset with high-quality, hand-labeled judgements. It facilitated the first comparison of systems on established corpora, tasks, and gold-labels (Schlechtweg et al. 2020).

However, despite being the largest and broadest existing evaluation framework, the definition of LSC used in Schlechtweg et al. (2020) – i.e., a binary classification and a ranking task – is a radical reduction of the full LSC task. The definition of LSC involves modelling of words and detecting (sense) changes, as well as generalising across many more time points, and disambiguating instances of words in the text. There cannot be only one universal model of a word: there are many ways to describe a word and its senses (see, for example, different dictionary definitions of the same word). So how do we devise evaluation data and methods such that different ways of defining meaning are taken into consideration when evaluating? Should a future evaluation dataset involve dividing the original sentences where a word is used in a particular sense into clusters with sentences that contributed to each sense, to avoid having to evaluate the different representations modelled for a word? How do we handle the uncertainty of which sense led to another? And how many new instances of change are needed to constitute semantic change?

Current work in unsupervised LSC is primarily limited to binary decisions of "change" or "no change" for each word. However, some go beyond the binary to include graded change (although these changes are then often used in figures, for illustrative purposes), and a possible classification of change type. Future work in LSC needs to include a discussion of what role the modelling of sense and signal should play in the evaluation of semantic change: how large does the correspondence between the model and the "truth" for a model need to be, for the results to be deemed accurate? Should we be satisfied to see our methods performing well on follow-up (or downstream) tasks but failing to give proper semantic representation? Evaluation heavily depends on task definition – and thus on the principle of fitness for use.[18] In addition, to study LSC with different task definitions we need to have datasets that reflect these perspectives and make use of task-specific definitions of both meaning and change during evaluation.

[17] The work by Hu et al. (2019) uses *dated* entries of the *Oxford English Dictionary* and thus provides an exception.

[18] A concept originally from Joseph M. Juran, and thoroughly discussed in Boydens (1999).

4.1 Types of evaluations

In the following subsections, we tackle two types of evaluation typically employed for LSC. We first discuss evaluation on ground-truth data, then tackle the promising evaluation on artificially-induced LSC data and argue that both should be used in a complementary fashion.

4.1.1 Ground truth

An important part of evaluation is determining what to evaluate. For example, some studies perform quantitative evaluation of regularities in the vocabulary as a whole. Regardless of other potential evaluation strategies, all existing work (also) evaluates change detected for a small number of lexical items – typically words – in a qualitative manner. This is done in one of two ways: either (i) a set of predetermined words are used for which there is an expected pattern of change, or (ii) the (ranked) output of the investigated method or methods is evaluated. Both of these evaluation strategies have the same aim, but with different (dis)advantages, which we discuss below.

(i) This evaluation strategy consists of creating a pre-chosen test set and has the advantage of requiring less effort as it removes the need to conduct a new evaluation for each change made to parameters such as size of time bins, or preprocessing procedure. The downside is, however, that the evaluation does not allow for new, previously unseen examples.[19] The pre-chosen words can be positive examples (words known to have changed), or negative examples (words known to be stable). Evaluation on only one class of words, positive or negative, does not properly measure the performance of a method. Let us say that we have a method that always predicts change, and we only evaluate on words that have changed. Unless we also evaluate exactly how the word has changed, or when, the method will always be 100% accurate. The best indicator of a method's performance is its ability to separate between positive and negative examples, and hence any pre-chosen test set should consist of words from both classes. However, we also need a proper discussion of the proportion of positive and negative examples in the test set, as the most likely scenario in any given text is "no change".

[19] It is, of course, always possible to augment this "gold set" with new examples. Gold truth creation, though, is extremely costly both in time and money: Schlechtweg et al. (2020) report a total cost of EUR 20,000 (1,000 hours) for 37 English words, 48 in German, 40 in Latin, and 31 in Swedish.

(ii) Evaluating the output of the algorithm allows us to evaluate the performance of a method "in the wild" and truly study its behaviour. Unfortunately, this evaluation strategy requires new evaluation with each change either to the method or the data, as there potentially can be a completely new set of words to evaluate each time. The words to evaluate can be chosen on the basis of a predetermined measure of change (e.g., largest / smallest cosine angle between two consecutive time periods, i.e., the words that changed the most or least), or a set of randomly chosen words. Once a set of words is determined, the evaluation of each word is done in the same manner as for the pre-chosen test set.

The *accuracy* of the evaluation, regardless of strategy chosen, depends on the way we determine if and how a word has changed. The ground-truth must be constructed from the data (corpus) on which the methods are trained because existing dictionaries might list changes seen in the *language*, that might not be present in the corpus, or vice versa. Requiring a method to find change that is not present in the underlying text, or considering detected changes as false because they are not present in a general-purpose dictionary, both lead to artificially low performance of the method. When (manually) creating ground-truth data for evaluation, sample sentences from the dataset should be read and taken into consideration, thus grounding the change in the dataset.

4.1.2 Simulated LSC

Obtaining ground-truth data for LSC is a difficult task as it requires skilled annotators and takes time to produce. The problem is exacerbated as the time depth of the language change phenomena increases and the languages at hand become rarer. This fact leads to a further requirement: expert annotators. The notion of "expert annotator" is problematic when judging senses in the past. Previous studies (e.g. Schlechtweg et al. 2018) note that historical linguists tend to have better inter-annotator agreement between themselves than with "untrained" native speakers – hinting at the fact that this is a skill that can be honed. The difficulty of engaging sufficiently many expert annotators is also a theoretical argument in favour of synthetic evaluation frameworks as a complement. In addition, some types of LSC are less frequent than others,[20] therefore requiring large amounts of text to be annotated in order to find enough samples. To alleviate these problems, simulating LSC in existing corpora has been suggested.

[20] Assuming that semantic change is power-law distributed, like most linguistic phenomena.

Simulating LSC is based on a decades-old procedure of inducing polysemy to evaluate word sense disambiguation systems (Gale et al. 1992, Schütze 1998). In this approach two or more words (e.g., *chair* and *sky*) are collapsed into a single word form (*chairsky*), thus conflating their meanings and creating a pseudo polysemous word (the original pair is removed from the lexicon). From another perspective, if this procedure unfolds through time (i.e., a word either gained or lost senses), then it can be considered to simulate LSC via changes to the number of senses of words. Indeed, this approach has been used extensively to simulate LSC (Cook & Stevenson 2010, Kulkarni et al. 2015, Rosenfeld & Erk 2018, Shoemark et al. 2019). However, the LSC that is induced in this way is rather synthetic, because it collapses unrelated words into a single word form, as opposed to the general view that finds the different senses to be semantically related (Fillmore & Atkins 2000). In order to provide a more accurate LSC simulation, Dubossarsky et al. (2019) accounted for the similarity of candidate words prior to their collapse, both creating related pairs (e.g., *chair* and *stool*) that better reflect true polysemy, and comparing the pair with the original approach of unrelated pairs (*chair* and *sky*). Schlechtweg & Schulte im Walde (2020) use SemCor, a sense-tagged corpus of English, to control for the specific senses each word has at each time point, thus providing an even more ecological model for simulated LSC.

The simulated approach to LSC has the potential to circumvent any bottleneck related to the need for annotators, and thus reduces costs. In addition, with careful planning, it should be possible to simulate any desirable type of LSC, regardless of its rarity in natural texts. As an added bonus, and certainly of interest to lexicographers, such an evaluation allows us to compute recall. In this scenario, recall would be proportional to the number of changed words that a given method can find. Using a synthetic change dataset is currently the only realistic scenario for determining the recall of our models and therefore, detecting how much change a method is able to capture. At the same time, it is hard to argue against the legitimate concern that these LSCs are artificial, and as such may not be the optimal way to evaluate detection by computational models. Certainly, synthetic change datasets are not optimal to study the natural linguistic phenomenon of semantic change, at least before we have a full understanding of the large-scale phenomena that we wish to study at which point we might no longer be in need for synthetic datasets. However, without the considerable effort to annotate full datasets, we are bound to use synthetic change evaluation sets – despite the inherent limitation described above. As a result, an important factor for future research becomes the creation of synthetic datasets that reflect the complex and varying nature of real language and real semantic change.

We stipulate that simulated datasets should be used alongside ground-truth testing, both with respect to pre-chosen test sets, as well as evaluating the output, to properly evaluate the ability of any method to detect LSC.

4.2 Quantifying meaning and meaning change

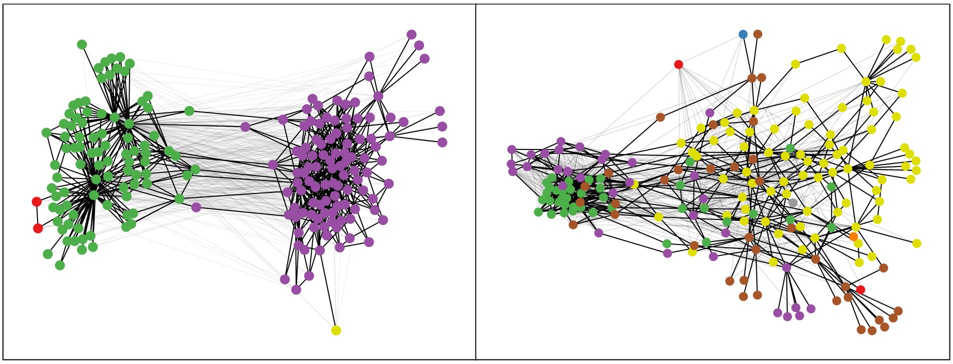


Figure 11.1: Word usage graphs of German *zersetzen* (left) and *Abgesang* (right).

To provide high-quality, ground-truth data for LSC where word meaning and change is grounded in a given corpus, we must perform manual annotation. However, first, we need to choose the relevant level of meaning so that we can quantify meaning distinctions and the change for a word based on the annotation. Recently, SemEval-2020 Task 1 (Schlechtweg et al. 2020) implemented a *binary* and a *graded* notion of LSC (Schlechtweg & Schulte im Walde 2020) in the shared task, which was partly adopted by a follow-up task on Italian (Basile et al. 2020). The annotation used typical meaning distinctions from historical linguistics (Blank 1997). Although the authors avoided the use of discrete word senses in the annotation by using graded semantic relatedness judgements (Erk et al. 2013, Schlechtweg et al. 2018), they grouped word uses post-hoc into hard clusters and interpreted all uses in a cluster as having the same sense. While this discrete view can work well in practice for some words (Hovy et al. 2006), it is inadequate for others (Kilgarriff 1997, McCarthy et al. 2016). In order to see this, consider Figure 11.1, showing the annotated and clustered uses for two words from the SemEval dataset: the uses of the word *zersetzen* on the left can clearly be partitioned into two main clusters, while the ones of *Abgesang* on the right have a less clearly clusterable structure.

A graded notion of meaning and change can be used to avoid having to cluster cases like the latter, though it is still unclear what the practical applications could be for LSC without discrete senses. The advantage of discrete word senses is that, despite their inadequacy for certain words, they are a widely used concept and also build a bridge to historical linguistics (Blank 1997, Blank & Koch 1999). This bridge is an important one, because the most straightforward application of LSC detection methods is for historical linguistics or lexicography. Nonetheless, there might be many task-specific definitions of LSC that could do without sense distinctions, and the issue is an interesting avenue for future work.

5 Related fields and applications

The field of LSC has close ties with two types of disciplines: those that study (i) *language*, and those that study (ii) *human activities*. In this section, we shed light on prominent work in these fields without claiming to be exhaustive, and discuss the potential of interactions with these fields.

5.1 Studying language

A great deal of existing work has gone into the study of language. Lexicography benefits a great deal from semantic representation in time, with works by, among others, Lau et al. (2012), Falk et al. (2014), Fišer & Ljubešić (2018), Klosa & Lüngen (2018), and Torres-Rivera & Torres-Moreno (2020). In this strand, methods for LSC can prove efficient at updating historical dictionaries: by using LSC approaches on large-scale corpora, it becomes possible to verify, at the very least, whether a sense was actually used *before* its current date in the dictionary. Senses cannot be post-dated, on the other hand; their absence from a corpus does not necessarily mean they did not exist elsewhere. Lexicographers can ideally use these methods to generate *candidates* for semantic change which would then be manually checked. They could also use sense-frequency data to paint the prominence of a word's senses through time, or even incorporate a quantified measure of similarity between senses of the same word – features that could also be incorporated in contemporary dictionaries.

Another strand, despite most work focusing solely on English, concerns language in general. In the past few years, there have been several attempts at testing hypotheses for laws of change which were proposed more than a century ago, or devising new ones. Xu & Kemp (2015) focus on two incompatible hypotheses: Bréal (1897)'s LAW OF DIFFERENTIATION (where near-synonyms are set to diverge across time) and Stern (1921)'s LAW OF PARALLEL CHANGE (where words sharing related meanings tend to move semantically in the same way). They showed quantitatively for English, in the Google Books corpus, that Stern's law of parallel change seems to be more rooted in evidence than Bréal's law of differentiation. Dubossarsky et al. (2015) ground their LAW OF PROTOTYPICALITY on the hypothesis of Geeraerts (1997) that a word's relation to the core prototypical meaning of its semantic category is crucial with respect to diachronic semantic change, and show using English data that prototypicality is negatively correlated with semantic change. Eger & Mehler (2016) postulate and show that semantic change tends to behave linearly in English, German and Latin. Perhaps the best-known example of such work within NLP, and often the only one cited, are the two laws of Hamilton et al. (2016b): CONFORMITY (stating that frequency is negatively correlated with semantic change), and INNOVATION (hypothesising that polysemy is positively correlated with semantic change).

Interestingly, since the NLP work above derives from observations that are replicable, quantitative, somewhat evidentiary, and not from a limited set of examples as was the case in the early non-computational days of semantic change research, previous laws elicited from quantitative investigations can be revisited. Such was the aim of Dubossarsky et al. (2017). They show that three previous laws (the law of prototypicality of Dubossarsky et al. 2015 and the laws of innovation and conformity by Hamilton et al. 2016b) are a byproduct of a confounding variable in the data, namely frequency, and are thus refuted. The paper calls for more stringent standards of proof when articulating new laws – in other words, robust evaluation.

As regards future work, we envision the field of LSC moving towards better use of linguistic knowledge. Traditional semantics and semantic change research is deeply rooted in theories that can now be computationally operationalized. Additionally, advances in computational typology and cross-lingual methods allow language change to be modelled for several similar languages at the same time (as started by Uban et al. 2019 and Frossard et al. 2020, for example), and to take into account theories of language contact. Other linguistic features can also be taken into account, and we hope to see more work going beyond "simple" lexical semantics.[21] The overview and discussions in this chapter have primarily targeted semantic change, often referred to as semasiological change in linguistic literature, while onomasiological change relates to different words used for the same concepts at different points in time. This general concept is often referred to as lexical replacement (Tahmasebi et al. 2018). Future work should attempt to resolve onomasiological and semasiological change in an iterative manner to ensure coherency in our models.

[21] An excellent example of this move forward can be seen in Fonteyn (2020), for example.

5.2 Studying human society

Along with the study of language itself, NLP techniques can be repurposed to serve different goals. With NLP methods maturing and technical solutions being made available to virtually anyone,[22] theories in other fields can be tested quantitatively. Quite obviously, since language is humans' best tool of communication, advanced techniques that tackle language are useful in many other fields, where they are often applied as-is, and sometimes modified to serve a different purpose. What is often disregarded in NLP, however, is what we need from those tangential disciplines in order to arrive at reliable models. One obvious answer to this question pertains to data resources. Those who are working on semantic change computation are heavily dependent on the data at their disposal, and should pay more attention to the type, diversity and quality of data they are working with, as discussed in Section 2.

In this subsection we focus on a few examples of how related fields have borrowed methods from LSC by using some examples from the literature, and attempt to give broad avenues for a continued mutualistic relationship between LSC and those fields.

A great deal of past human knowledge that has survived is stored in texts. Historical research[23] is arguably a large beneficiary of proper semantic representations of words in time: an often voiced critique in historical scholarship relates to chronological inconsistencies – anachronisms (Syrjämäki 2011). As reported by Zosa et al. (2020), Hobsbawm (2011) stated that "the most usual ideological abuse of history is based on anachronism rather than lies". This fact leads to many historians trying to "see things their [the people of the past's] way" (Skinner 2002). Somewhat similarly, Koselleck (2010) underlines the "veto right of the sources". However, for one to use the sources properly, they need to be understood correctly, and proper modelling of a word's semantics across time can definitely help historians interpret past events. Furthermore, the "concepts as factors and indicators of historical change" of Koselleck (2004: 80) highlights the importance of language as a window on the past. There is a growing body of work with quantitative diachronic text mining (such as word embeddings and (dynamic) topic models) within humanities research which clearly benefits from NLP methods, but can similarly inform LSC. For example, Heuser (2017)[24] studies the difference

[22] For example, extremely large-scale pretrained models are shared on platforms such as Hugging Face (https://huggingface.co/models) allowing anyone to download and use them with limited hardware; while efficient libraries such as gensim (Řehůřek & Sojka 2010) make the training of type embeddings possible on personal laptops.

[23] For clarity's sake, we do not differentiate between "historical research", "digital history", "computational history", and "digital humanities". For a broader discussion about field-naming in the (digital) humanities, refer to Piotrowski (2020).

[24] See https://twitter.com/quadrismegistus/status/846105045238112256 for a visualisation.

between abstract and concrete words in different literary subgenres. Similarly, Björck and co-authors[25] study the Swedish word for 'market', *marknad*, and describe a change in abstractness through time: from a physical market (as a noun), to more and more abstract notions (such as 'labour market') and even to the point where the noun is used as a modifier (e.g. 'market economy'). These observations teach us not only about the word itself, but also about the world. If words such as *table* or *car* are relatively straightforward to define and probably easier to model (see e.g. Reilly & Desai 2017 who show that concrete words tend to have denser semantic neighbourhoods than abstract words), what lessons can we learn from such work when representing abstract concepts? LSC methods should strive to include such key information in its methods.

Claiming that current LSC methods can "solve historical research"[26] and provide definitive answers to long-studied phenomena would be, at best, extremely misleading. Indeed, while LSC methods can model a word's sense(s) across time, humanists (or political scientists, for that matter) can be described as studying *concepts*. An emerging or evolving concept, almost by definition, will not be constrained to a single word. Rather, methods will probably have to be adapted to study a cluster of words[27] – either manually chosen (Kenter et al. 2015, Recchia et al. 2016), or selected in a more data-driven way (Tahmasebi 2013, Hengchen et al. to appear). These clusters will be the basis for historical contextualisation and interpretation. The same ad-hoc adaptation is to be found in political science: a recent example of NLP methods making their way in (quantitative) political science is the work of Rodman (2020) where the author fits both an LDA model on more than a century of newspapers as well as a supervised topic model – using 400 hand-annotated documents by several annotators with a high inter-annotator agreement – so as to produce a gold standard to evaluate diachronic word embeddings, with the final aim of studying the evolution of *concepts* such as 'gender' and 'race'. Similar work is undertaken by Indukaev (2021), who studies modernisation in Russia and convincingly describes the benefits and limitations of topic models and word embeddings for such a study.

While extremely promising, our current methods fail to serve related fields that would benefit greatly from them: as of now, most LSC approaches simply model words, and not concepts – again underlining the need for task-specific meaning tackled in Section 3.

[25] Presentation by Henrik Björck, Claes Ohlsson, and Leif Runefelt given at the Workshop on automatic detection of language change 2018 co-located with SLTC 2018, Stockholm. For more details, see Ohlsson (2020).

[26] Or any field concerned with diachronic textual data.

[27] These clusters of words are related to what linguists call lexical fields, a term that in our experience is not widely used in other disciplines.

6 Conclusions

In this chapter, we have outlined the existing challenges in reconstructing the semantic change evident in large diachronic corpora.

Currently, as was made obvious in Section 2, the field suffers from several limitations when it comes to data. Indeed, we believe that future work should strive to use, and produce, high-quality text in many languages and different genres. This point is crucial: if we are to detect semantic change, our data needs to have certain precise qualities, as well as well-defined metadata. It is thus difficult for LSC researchers to rely on data created for other NLP fields. As a plea to the larger community, we count on the field not to make the mistake of assuming that the available textual data is representative of the language at hand. We further hope that in the future, meaning can be modelled by using not only text, but also multi-modal data.

Modelling is notoriously difficult, but, to paraphrase Box (1976), models being inherently wrong does not ineluctably make them useless. A crucial component to the useful modelling of meaning and of change outlined in Section 3 is the definition of what meaning is. Whether we draw from the large body of work that exists in traditional semantics research or start from scratch with a new definition of what *computational* meaning is, we hope researchers in our field can come together and agree on *what* is and should be modelled. Only with shared, solid models of meaning can the field move forward with the complexity, possibly intractable, of modelling meaning change. A word's semantics have changed – but how?

Echoing the complexity of modelling information from data is the consistency needed in the evaluation of a model's output. Section 4 makes the point that without a homogeneous, somewhat large-scale evaluation framework across languages such as the one proposed in Schlechtweg et al. (2020), researchers cannot confidently rely on conclusions from previous work to move forward. Since ground-truth creation is expensive both in time and money and is ineluctably limited to a single corpus, we encourage the community to pay attention to synthetic evaluation techniques which have the potential to circumvent cost, evaluate different types of semantic change, and tackle different temporal granularities. Our field is rich in methods but in dire need of comparable results. This can be partially solved with robust, thorough, and shared evaluation practices.

Being able to model and detect different types of semantic change is important in LSC, and also in related disciplines such as lexicography and historical linguistics. The history of ideas, and any area concerned with the diachronic study of textual data, would greatly benefit from our methods – if they are robust. In addition, we believe that there is potential for a mutualistic relationship with those parallel fields not only contributing theory or domain expertise but also echoing the need for the proper modelling of words, senses, and types of change.

Author contributions

All authors contributed equally, and the ordering is determined in a round robin fashion.

Acknowledgements

We thank our reviewers for their very helpful comments. This chapter benefited from insightful comments from Lars Borin. This work has been funded in part by the project *Towards computational lexical semantic change detection* supported by the Swedish Research Council (2019–2022; dnr 2018-01184), and *Nationella språkbanken* (the Swedish National Language Bank), jointly funded by the Swedish Research Council (2018–2024; dnr 2017-00626) and the ten partner institutions. Dominik Schlechtweg was supported by the Konrad Adenauer Foundation and the CRETA Centre funded by the German Ministry for Education and Research (BMBF) during the writing of this chapter. Haim Dubossarsky is supported by the Blavatnik Postdoctoral Fellowship Programme. The authors wish to thank the Department of Swedish at the University of Gothenburg for providing financial support for language editing services.

Abbreviations

ACL	Association for Computational Linguistics
COHA	corpus of historical American English
LDA	latent Dirichlet allocation
LSC	lexical semantic change
NLP	natural language processing
OCR	Optical Character Recognition
SLTC	Swedish Language Technology Conference
WSD	word sense disambiguation

References

Aristotle. 1898. *Poetics.* (English translation by Ingram Bywater). Oxford: The Clarendon Press.

Bamler, Robert & Stephan Mandt. 2017. Dynamic word embeddings. In Doina Precup & Yee Whye Teh (eds.), *Proceedings of the 34th international conference on machine learning* (Proceedings of Machine Learning Research 70), 380–389. Sydney: PMLR.

Basile, Pierpaolo, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti & Rossella Varvara. 2020. Overview of the EVALITA 2020 diachronic lexical semantics (DIACR-ita) task. In Valerio Basile, Danilo Croce, Maria Di Maro & Lucia C. Passaro (eds.), *Proceedings of the 7th evaluation campaign of natural language processing and speech tools for Italian (EVALITA 2020).*

Bender, Emily M. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6(3). 1–26.

Bender, Emily M. & Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL 2020*, 5185–5198. Online: ACL. DOI: 10.18653/v1/2020.acl-main.463.

Blank, Andreas. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen.* Tübingen: Niemeyer.

Blank, Andreas. 1999. Why do new meanings occur? A cognitive typology of the motivations for lexical semantic change. In Andreas Blank & Peter Koch (eds.), *Historical semantics and cognition* (Cognitive Linguistics Research 13), 61–90.

Blank, Andreas & Peter Koch. 1999. *Historical semantics and cognition.* Berlin: Walter de Gruyter.

Bloomfield, Leonard. 1933. *Language.* New York: Henry Holt.

Bowern, Claire. 2019. Semantic change and semantic stability: Variation is key. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 48–55.

Box, George EP. 1976. Science and statistics. *Journal of the American Statistical Association* 71(356). 791–799.

Boydens, Isabelle. 1999. *Informatique, normes et temps.* Bruxelles: Bruylant.

Bréal, Michel. 1897. *Essai de sémantique.* Paris: Hachette.

Bruni, Elia, Gemma Boleda, Marco Baroni & Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (Volume 1: Long papers)*, 136–145.

Cook, Paul, Jey Han Lau, Diana McCarthy & Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014: Technical papers*, 1624–1635. Dublin: ACL. https://www.aclweb.org/anthology/C14-1154.

Cook, Paul & S. Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In N. C. C. Chair, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.), *Proceedings of LREC 2010*. Valletta: ELRA.

Davies, Mark. 2002. *The corpus of historical American English (COHA): 400 million words, 1810-2009*. Provo: Brigham Young University.

Denny, Matthew J. & Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis* 26(2). 168–189.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi & Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of ACL 2019*, 457–470. Florence: ACL. DOI: 10.18653/v1/P19-1044.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer & Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *Proceedings of NetWordS 2015* (CEUR Workshop Proceedings 1347), 66–70. Pisa.

Dubossarsky, Haim, Daphna Weinshall & Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of EMNLP 2017*, 1136–1145. Copenhagen: ACL. DOI: 10.18653/v1/D17-1118.

Eger, Steffen & Alexander Mehler. 2016. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of ACL 2016 (Volume 2: Short papers)*, 52–58. Berlin: ACL. DOI: 10.18653/v1/P16-2009.

Erk, Katrin, Diana McCarthy & Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics* 39(3). 511–554.

Falk, Ingrid, Delphine Bernhard & Christophe Gérard. 2014. From non word to new word: Automatically identifying neologisms in French newspapers. In *Proceedings of LREC 2014*, 4337–4344. Reykjavik: ELRA.

Fillmore, Charles J. & Beryl T. S. Atkins. 2000. Describing polysemy: The case of 'crawl'. *Polysemy: Theoretical and computational approaches* 91. 110.

Fišer, Darja & Nikola Ljubešić. 2018. Distributional modelling for semantic shift detection. *International Journal of Lexicography* 32(2). 1–21. DOI: 10.1093/ijl/ecy011.

Fonteyn, Lauren. 2020. What about grammar? Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions. *CEUR Workshop Proceedings* 1613. http://ceur-ws.org/Vol-2723/short15.pdf.

Frossard, Esteban, Mickael Coustaty, Antoine Doucet, Adam Jatowt & Simon Hengchen. 2020. Dataset for temporal analysis of English-French cognates. In *Proceedings of LREC 2020*, 855–859. Marseille: ELRA. https://www.aclweb.org/anthology/2020.lrec-1.107.

Gale, William A., K. W. Church & D. Yarowsky. 1992. Work on statistical methods for word sense disambiguation. In *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 23–25. http://www.aaai.org/Papers/Symposia/Fall/1992/FS-92-04/FS92-04-008.pdf.

Geeraerts, Dirk. 1997. *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford: Oxford University Press.

Giulianelli, Mario, Marco Del Tredici & Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of ACL 2020*, 3960–3973. Online: ACL. DOI: 10.18653/v1/2020.acl-main.365.

Gulordava, Kristina & Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, 67–71. Edinburgh: ACL. https://www.aclweb.org/anthology/W11-2508.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016a. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of EMNLP 2016*, 2116–2121. Austin: ACL. DOI: 10.18653/v1/D16-1229.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of ACL 2016 (Volume 1: Long papers)*, 1489–1501. Berlin: ACL. DOI: 10.18653/v1/P16-1141.

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2-3). 146–162.

Hätty, Anna, Dominik Schlechtweg & Sabine Schulte im Walde. 2019. SURel: A gold Standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, 1–8. Minneapolis.

Hellrich, Johannes & Udo Hahn. 2016. Bad company: Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016: Technical papers*, 2785–2796. Osaka: ACL. https://www.aclweb.org/anthology/C16-1262.

Hengchen, Simon. 2017. *When does it mean? Detecting semantic change in historical texts*. Brussels: Université libre de Bruxelles. (Doctoral dissertation).

Hengchen, Simon, Ruben Ros & Jani Marjanen. 2019. A data-driven approach to the changing vocabulary of the *nation* in English, Dutch, Swedish and Finnish newspapers, 1750–1950. In *Proceedings of the Digital Humanities (DH) conference 2019*.

Hengchen, Simon, Ruben Ros, Jani Marjanen & Mikko Tolonen. to appear. A data-driven approach to studying changing vocabularies in historical newspaper collections. *Digital Scholarship in the Humanities*. DOI: 10.1093/llc/fqab032.

Heuser, Ryan James. 2017. Word vectors in the eighteenth century. In *Book of abstracts of the 2017 Digital Humanities conference (DH2017)*. Montréal.

Hobsbawm, Eric. 2011. *On History*. London: Hachette UK.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of NAACL-HLT: Short papers* (NAACL-Short '06), 57–60. New York: ACL.

Hu, Renfen, Shen Li & Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of ACL 2019*, 3899–3908. Florence: ACL. DOI: 10.18653/v1/P19-1379.

Indukaev, Andrey. 2021. Studying ideational change in Russian politics with topic models and word embeddings. In Daria Gritsenko, Mariëlle Wijermars & Mikhail Kopotev (eds.), *Palgrave handbook of Digital Russia Studies*, 443–465. Basingstoke: Palgrave Macmillan.

Jawahar, Ganesh & Djamé Seddah. 2019. Contextualized diachronic word representations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 35–47. Florence: ACL.

Kaiser, Jens, Dominik Schlechtweg, Sean Papay & Sabine Schulte im Walde. 2020. IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in lexical semantic change detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*. Barcelona: ACL.

Kenter, Tom, Melvin Wevers, Pim Huijnen & Maarten De Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1191–1200.

Kilgarriff, Adam. 1997. "I don't believe in word senses". *Computers and the Humanities* 31(2). 91–113.

Kim, Yoon, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde & Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, 61–65. Baltimore: ACL. DOI: 10.3115/v1/W14-2517.

Klosa, Annette & Harald Lüngen. 2018. New German Words: Detection and Description. In *Proceedings of the XVIII EURALEX international congress lexicography in global contexts, 17–21 July 2018, Ljubljana*, 559–569. Ljubliana: Ljubljana University Press.

Koplenig, Alexander. 2016. *Analyzing lexical change in diachronic corpora*. Universität Mannheim. (Doctoral dissertation). http://nbn-resolving.de/urn:nbn:de:bsz:mh39-48905.

Koselleck, Reinhart. 2004. *Futures past: On the semantics of historical time*. New York, NY: Columbia University Press.

Koselleck, Reinhart. 2010. *Vom Sinn und Unsinn der Geschichte: Aufsätze und Vorträge aus vier Jahrzehnten*. Carsten Dutt (ed.). 1st edn. Berlin: Suhrkamp.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi & Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on the World Wide Web*, 625–635. Florence: ACM. DOI: 10.1145/2736277.2741627.

Kutuzov, Andrey, Erik Velldal & Lilja Øvrelid. 2017. Temporal dynamics of semantic relations in word embeddings: An application to predicting armed conflict participants. In *Proceedings of EMNLP 2017*, 1824–1829. Copenhagen: ACL. DOI: 10.18653/v1/D17-1194.

Lau, Jey Han, Paul Cook, Diana McCarthy, David Newman & Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of EACL 2012*, 591–601. Avignon: ACL. https://www.aclweb.org/anthology/E12-1060.

Magué, Jean-Philippe. 2005. *Changements sémantiques et cognition: Différentes méthodes pour différentes échelles temporelles*. Lyon: Université Lumière. (Doctoral dissertation).

McCarthy, Diana, Maria Apidianaki & Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics* 42(2). 245–275.

McGillivray, Barbara, Simon Hengchen, Viivi Lähteenoja, Marco Palma & Alessandro Vatri. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities* 34(4). 893–907.

Meillet, Antoine. 1905. Comment les mots changent de sens. *Année Sociologique*.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331. 176–182.

Mitra, Sunny, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal & Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering* 21(5). 773–798.

Ohlsson, Claes. 2020. Market language over time. Combining corpus linguistics and historical discourse analysis in a study of market in Swedish press texts.

In Joacim Hansson & Jonas Svensson (eds.), *Doing digital humanities: Concepts, approaches, cases*, vol. 1, 199–218. Växjö: Linnaeus University.

Pechenick, Eitan Adam, Christopher M. Danforth & Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE* 10(10). e0137041.

Perrone, Valerio, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith & Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 56–66. Florence: ACL. DOI: 10.18653/v1/W19-4707.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018: Volume 1 (Long papers)*, 2227–2237. New Orleans: ACL. DOI: 10.18653/v1/N18-1202.

Piotrowski, Michael. 2012. Natural language processing for historical texts. *Synthesis lectures on human language technologies* 5(2). 1–157.

Piotrowski, Michael. 2020. Ain't no way around it: Why we need to be clear about what we mean by "digital humanities". *SocArXiv*. DOI: 10.31235/osf.io/d2kb6.

Ponti, Edoardo Maria, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova & Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics* 45(3). 559–601. DOI: 10.1162/coli_a_00357.

Prescott, Andrew. 2018. Searching for Dr. Johnson: The digitisation of the Burney Newspaper Collection. In Siv Gøril Brandtzæg, Paul Goring & Christine Watson (eds.), *Travelling chronicles: News and newspapers from the early modern period to the eighteenth century*, 51–71. Leiden: Brill.

Recchia, Gabriel, Ewan Jones, Paul Nulty, John Regan & Peter de Bolla. 2016. Tracing shifting conceptual vocabularies through time. In *European knowledge acquisition workshop*, 19–28. Springer.

Řehůřek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, 45–50. http://is.muni.cz/publication/884893/en. Valletta: ELRA.

Reilly, Megan & Rutvik H. Desai. 2017. Effects of semantic neighborhood density in abstract and concrete words. *Cognition* 169. 46–53. DOI: 10.1016/j.cognition.2017.08.004.

Reisig, Karl. 1839. *Vorlesungen über lateinische Sprachwissenschaft*. Leipzig: Lehnhold.

Rodda, Martina A., Marco S.G. Senaldi & Alessandro Lenci. 2017. Panta rei: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics* 3(1). 11–24. https://arpi.unipi.it/handle/11568/891899.

Rodman, Emma. 2020. A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis* 28(1). 87–111.

Rosenfeld, Alex & Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of NAACL-HLT 2018: Volume 1 (Long papers)*, 474–484. New Orleans: ACL. DOI: 10.18653/v1/N18-1044.

Rudolph, Maja R. & David M. Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of WWW 2018*, 1003–1011. ACM. DOI: 10.1145/3178876.3185999.

Säily, Tanja. 2016. Sociolinguistic variation in morphological productivity in eighteenth-century English. *Corpus Linguistics and Linguistic Theory* 12(1). 129–151.

Schlechtweg, Dominik, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde & Daniel Hole. 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of CoNLL 2017*, 354–367. Vancouver: ACL.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky & Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of SemEval 2020*, 1–23. Barcelona: ACL. https://www.aclweb.org/anthology/2020.semeval-1.1.

Schlechtweg, Dominik & Sabine Schulte im Walde. 2020. Simulating lexical semantic change from sense-annotated data. In C. Cuskley, M. Flaherty, H. Little, Luke McCrohon, A. Ravignani & T. Verhoef (eds.), *The Evolution of Language: Proceedings of the 13th International Conference (EVOLANGXIII)*.

Schlechtweg, Dominik, Sabine Schulte im Walde & Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of NAACL 2018*. ACL.

Schofield, Alexandra, Måns Magnusson & David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of EACL 2017 (Volume 2: Short papers)*, 432–436.

Schofield, Alexandra & David Mimno. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the ACL* 4. 287–300. DOI: 10.1162/tacl_a_00099.

Schofield, Alexandra, Laure Thompson & David Mimno. 2017. Quantifying the effects of text duplication on semantic models. In *Proceedings of EMNLP 2017*, 2737–2747. Copenhagen: ACL. DOI: 10.18653/v1/D17-1290.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–123.

Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale & Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of EMNLP-IJCNLP 2019*, 66–76. Hong Kong: ACL.

Skinner, Quentin. 2002. *Visions of politics. Vol. 1, Regarding method.* Cambridge: Cambridge University Press.

Stern, Nils Gustaf. 1921. *Swift, swiftly and their synonyms. A contribution to semantic analysis and theory.* Gothenburg: Göteborgs universitet. (Doctoral dissertation).

Syrjämäki, Sami. 2011. *Sins of a historian: Perspectives on the problem of anachronism.* Tampere: Tampere University Press.

Tahmasebi, Nina. 2013. *Models and algorithms for automatic detection of language evolution.* Gottfried Wilhelm Leibniz Universität Hannover. (Doctoral dissertation). http://edok01.tib.uni-hannover.de/edoks/e01dh13/771705034.pdf.

Tahmasebi, Nina, Lars Borin & Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint 1811.06278.*

Tahmasebi, Nina, Lars Borin, Adam Jatowt & Yang Xu (eds.). 2019. *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change.* Florence: ACL. https://www.aclweb.org/anthology/W19-4700.

Tahmasebi, Nina & Simon Hengchen. 2019. The strengths and pitfalls of large-scale text mining for literary studies. *Samlaren: Tidskrift för svensk litteraturvetenskaplig forskning* 140. 198–227.

Tahmasebi, Nina, Kai Niklas, Gideon Zenz & Thomas Risse. 2013. On the applicability of word sense discrimination on 201 years of modern English. *International Journal on Digital Libraries* 13. 135–153. DOI: 10.1007/s00799-013-0105-8.

Tahmasebi, Nina & Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of RANLP 2017*, 741–749. Varna: INCOMA Ltd. DOI: 10.26615/978-954-452-049-6_095.

Tangherlini, Timothy R. & Peter Leonard. 2013. Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics* 41(6). 725–749.

Thompson, Laure & David Mimno. 2018. Authorless topic models: Biasing models away from known structure. In *Proceedings of COLING 2018*, 3903–3914. Santa Fe: ACL. https://www.aclweb.org/anthology/C18-1329.

Torres-Rivera, Andrés & Juan-Manuel Torres-Moreno. 2020. Detecting new word meanings: A comparison of word embedding models in Spanish. *arXiv preprint 2001.05285*.

Traugott, Elizabeth Closs & Richard B. Dasher. 2001. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Tsakalidis, Adam & Maria Liakata. 2020. Sequential modelling of the evolution of word representations for semantic change detection. In *Proceedings of EMNLP 2020*, 8485–8497. Online: ACL. https://www.aclweb.org/anthology/2020.emnlp-main.682.

Uban, Ana, Alina Maria Ciobanu & Liviu P. Dinu. 2019. Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 161–166. Florence: ACL. DOI: 10.18653/v1/W19-4720.

Voigt, Rob, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky & Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* 114(25). 6521–6526. DOI: 10.1073/pnas.1702413114.

Wijaya, Derry Tanti & Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of DETECT 2011*, 35–40. ACM.

Xu, Yang & Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual meeting of the Cognitive Science Society, CogSci 2015*. Pasadena.

Zosa, Elaine, Simon Hengchen, Jani Marjanen, Lidia Pivovarova & Mikko Tolonen. 2020. Disappearing discourses: Avoiding anachronisms and teleology with data-driven methods in studying digital newspaper collections. In *Book of abstracts of the 2020 digital Humanities in the Nordic countries (DHN) conference*.

Name index

Computational approaches to semantic change

Semantic change – how the meanings of words change over time – has preoccupied scholars since well before modern linguistics emerged in the late 19th and early 20th century, ushering in a new methodological turn in the study of language change. Compared to changes in sound and grammar, semantic change is the least understood. Ever since, the study of semantic change has progressed steadily, accumulating a vast store of knowledge for over a century, encompassing many languages and language families.

Historical linguists also early on realized the potential of computers as research tools, with papers at the very first international conferences in computational linguistics in the 1960s. Such computational studies still tended to be small-scale, method-oriented, and qualitative. However, recent years have witnessed a sea-change in this regard. Big-data empirical quantitative investigations are now coming to the forefront, enabled by enormous advances in storage capability and processing power. Diachronic corpora have grown beyond imagination, defying exploration by traditional manual qualitative methods, and language technology has become increasingly data-driven and semantics-oriented. These developments present a golden opportunity for the empirical study of semantic change over both long and short time spans.

A major challenge presently is to integrate the hard-earned knowledge and expertise of traditional historical linguistics with cutting-edge methodology explored primarily in computational linguistics.

The idea for the present volume came out of a concrete response to this challenge. The *1st International Workshop on Computational Approaches to Historical Language Change* (LChange'19), at ACL 2019, brought together scholars from both fields.

This volume offers a survey of this exciting new direction in the study of semantic change, a discussion of the many remaining challenges that we face in pursuing it, and considerably updated and extended versions of a selection of the contributions to the LChange'19 workshop, addressing both more theoretical problems – e.g., discovery of "laws of semantic change" – and practical applications, such as information retrieval in longitudinal text archives.

ISBN 978-3-96110-312-6