MSc Artificial Intelligence
Master Thesis

Lexical Semantic Change Analysis with Contextualised Word Representations

Mario Giulianelli
11567252

36 ECTS
Defended on July 18, 2019

Supervisor
Dr Raquel Fernández

Assessor
Dr Lisa Beinborn

Co-supervisor
Marco Del Tredici

Institute for Logic, Language and Computation

University of Amsterdam
Institute for Advanced Study

Abstract

How does the meaning of a word change over time? This thesis introduces a bottom-up procedure that aggregates word usages into groups of meaningful usage types in order to detect and investigate lexical semantic change within diachronic collections of texts. Computational measures of semantic change have relied on distributional and predictive word representations as well as on models for word sense induction. While the first fail to take word polysemy into account, the second involve an a priori selection of the number of underlying word senses, use a limited context window and treat sentences as bags of words. In order to address these limitations, we use a neural language model to obtain contextualised word representations which are uniquely defined by a word form together with its entire sentential context, and we aggregate said representations by automatically selecting the number of saliently different usage types, in a data-driven fashion. We then propose three metrics of semantic shift to quantify the degree of change undergone by a word and evaluate them against human judgements. Furthermore, we analyse the linguistic properties that guide the formation of clusters of word usages and probe our method with various types of semantic change. Results show that we are able to detect, in corpora of varying temporal granularity, the narrowing, broadening, and metaphorisation of a word's interpretation. We can recognise cultural drifts driven by technological innovations, cultural transitions, and specific events, as well as more subtle linguistic shifts such as changes in the subcategorisation frames of nouns and verbs. Besides its empirical findings, this thesis demonstrates that language models and contextualised word representations constitute a versatile and fruitful framework for computational analyses of language change and variation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the fourteenth century the word *boy* used to refer to a male assistant, servant, slave, or to a male person born of humble parentage, whereas *girl* used to refer to a child or a young person of either sex (Oxford English Dictionary). By the fifteenth century a new, narrower usage had emerged for the word *girl*: in phrases such as *prety gyrle* it designated exclusively female individuals (OED). A century later, *boy* had lost its negative connotation and was more broadly used to refer to any male child, becoming the masculine counterpart of *girl*:

(1) *Whose child is that you beare so tenderly? Is it a* ***boy*** *or* ***girle****, I praie ye tell?* (OED)
(1594, R. Wilson Coblers Phrophesie 1.10180)

This example from Bybee (2015) shows that words generalise or specialise their meaning over the course of long time periods. Their meaning shifts due to internal linguistic processes as well as cultural factors like new technologies: the word *virtual* used to denote the property of almost being a particular thing or almost having a certain quality; nowadays it is also used in the sense of something which does not physically exists but appears to do so thanks to a computer simulation. Changes in meaning can also occur at a faster pace, e.g. to fulfil the communicative needs of specific speech communities. In 2017, a few months after the adoption of a fluorescent yellow football kit, Liverpool FC fans began to use the word *highlighter* to ironically refer to the new kit (Del Tredici et al., 2019).

Why is a word used in a new way? Why does an entity or concept cease to be designated by a word? Understanding language change helps us explain similarities and differences among languages of the world as well as variation within single languages. Perhaps more importantly, as language is inherently dynamic and continuously changing, understanding the processes that lead to change can provide us with more general insights into how language is used to interactively create meaning and to perform socially recognised actions.

With this goal in mind, we propose to model diachronic lexical change using not abstract representations of word forms or word senses but rather representations of unique contextualised word *usages*. For the operationalisation of this paradigm shift, we use a large pre-trained language model to obtain contextualised word representations (Devlin et al., 2019) and then cluster these representations into meaningful and interpretable agglomerates of word usages. By comparing temporally adjacent agglomerates and quantifying their differences, we are able to detect the narrowing, broadening, and metaphorisation of a word's interpretation. As for the nature of the detected change, our method can recognise cultural drifts, i.e. polysemisation processes driven by cultural and technological innovation as well as by specific events, and it identifies linguistic shifts such as changes in the subcategorisation frames of nouns and verbs.

Furthermore, our tracking procedure is applicable to language corpora that cover varying time scales and it is entirely data-driven: it does not require researchers to e.g. determine the number of underlying senses that a word is expected to have but rather it is able to induce interpretable types of word usages directly from the corpus. This produces results that are fine-grained with respect to the data set at hand. Indeed, by deploying our method on both historical and conversational corpora of English, we show that

it is particularly apt for the analysis of socially pervasive long-term shifts yet it can be fine-tuned to also recognise short-term community-specific changes in word meaning.

Finally, although it requires no linguistic annotation and little-to-no training, our approach is very versatile. Most experiments are performed in a zero-shot setting, thus they only require information about the time period where texts were written, and they suggest that the proposed techniques are not limited to the study of monolingual historical change, but can be used to investigate variation within a language as well as cross-linguistic patterns of variation and change.

1.1 Language as an ever-changing social instrument

For entirely unambiguous and felicitous communication, individuals would require a symbol for each action they need to perform. One symbolic unit, i.e. one word, would be used at the dining table to ask someone if they can pass the salt, another symbol would be used to denote salt as a mineral, and yet another, distinct symbol would be used to refer to a salt shaker. It is clearly unfeasible for speakers, however, to memorise symbol-function pairs that serve every possible need for reference, conceptualisation and interaction. Therefore, in order to codify a large variety of complex intentions, speakers rearrange a low amount of simple constituents, combining atomic symbols into larger symbolic units. This recursive construction of symbolic expressions allows for the productivity of language and is most often guided by the principle of compositionality: the meaning of a complex expression is determined by the meaning of its simple constituents and by the rules used to assemble them. Inevitably, this process is subject to human cognitive constraints: speakers must learn to use and recognise patterns of symbols, and to associate them with communicative functions.

There is another, perhaps more stringent constraint which stems from the fundamental human dimension of sociality. Not only must speakers ensure they remember all the mappings from forms to functions as isolated individuals, they also need to be able to use these mappings in the real world to perform actions and interactions (Searle, 1975, 1985). To this end interpretations of symbols, and of sequential combinations thereof, must be shared. Indeed speakers learn how to use language *together*, by interacting with each other and with their environment, and interpretations undergo processes of conventionalisation within groups of speakers in order to become shared (Milroy, 1992; Traugott and Trousdale, 2013). The participants and the social environment of a symbolic interaction are active forces in the processes of conventionalisation, and constitute what can be generically termed as *context*. As Searle and Austin, among others, have brought to light, all language use is contextual—situated according to precise social as well as spatio-temporal coordinates (Searle, 1975; Austin, 1975; Brugman, 1988). Context *shapes* linguistic interaction as it provides speakers and listeners with cues for the interpretation of symbols, and it *allows* interaction as although no two interactants share the exact same symbolic repertoire they can use words as situated "instructions to create meanings" (Traugott, 2017).

The social coordinates of context, however, are necessarily subject to variation. Most individuals experience this variation on a daily basis, e.g. when they move from their family environment to their work environment. As the range of social contexts is wide and ever-changing, a speech community must somehow determine a limited collection of symbols that can ideally satisfy communicative needs in any foreseeable context. Word polysemy is a necessary consequence of this selective process: it allows speakers to use the same word in different contexts to perform different actions. In other words, polysemy is not a static phenomenon: it is the result of a dynamic balancing activity between the maximisation of informativeness (ensuring that one's instructions to create meaning are correctly followed) and the minimisation of effort (limiting the cognitive overload that comes with remembering word forms and their possible functions) (Zipf, 1949; Ramscar and Baayen, 2013; Baayen et al., 2017). Hence, the senses of a word, i.e. the concepts associated with it (Traugott, 2017), can be acquired, lost, and they can shift over time. This dynamicity is observable e.g. in the variance of the type-sense ratio across languages, registers, and epochs.

The second source of instability for the symbolic mappings of which language consists is the variability of spatio-temporal coordinates. Different speech communities can use the same form to refer to different concepts and refer to the same concept with different forms. If they do so in the same time

period, we typically refer to this phenomenon as *variation*. Similarly, speech communities may use the same form with different goals in different epochs (and so do speakers across their lifespan (Baayen et al., 2017)). Permanent modifications of this kind constitute *change*. Far from being a disadvantage, this malleability makes language flexible enough to resist to the constant evolution of communicative needs. The alternative—an unfeasible one—would require designing a language *once* and in such a way that it is apt for conceptualisation and interaction concerning events and entities that are not yet known to the speech community.

1.2 Analysing language change

Both variation and change can in truth occur at all linguistic levels: speakers can vary the way words are pronounced, the way words can be formed from smaller units (and what even counts as a word), the way words are arranged together into phrases and periods, and the types of structures that can be built from periods themselves. Research on variation and change has indeed involved many linguistic variables, drawn from phonology, syntax, semantics, and discourse. For the analysis, in particular, of the diachronic dynamics of lexical semantics, recent approaches have largely focused on shift detection, the task of deciding whether and to what extent the concept evoked by a word has changed between adjacent time periods (e.g. Gulordava and Baroni, 2011; Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016; Bamler and Mandt, 2017; Rosenfeld and Erk, 2018). This line of work relies on distributional and predictive word representation models, therefore word types have been mostly used as unit of representation and unit of analysis. Such *type-based* models depend on a strong simplification—that one abstract representation is sufficient to model polysemous words. E.g. the same word representation of *highlighter* would be used in both of the following contexts[1]:

(2) a. *Apply the **highlighter** under the eyes, above the brows and on the browbone.*
b. *This **highlighter** does not bleed through paper and it does not smear ink across the page.*

This is why other researchers have directed their attention to word senses and their induction over time periods, typically detecting novel senses based on the diachronic divergence between sense distributions. Various Bayesian models have been developed for this task (Lau et al., 2012; Cook et al., 2014); the latest advances include the SCAN model (Frermann and Lapata, 2016) and dynamic embeddings (Rudolph and Blei, 2018). A non-Bayesian approach has been put forward by Mitra et al. (2014, 2015), who use dependency label features to define sense clusters.

It has been argued, however, that senses themselves are a discretisation of something that is continuous in nature and partially undetermined, and that words are modulated by speakers within each and every conversation to convey a contingently intended interpretation (Brugman, 1988; Paradis, 2011; Ludlow, 2014). As an example, the occurrence of *highlighter* in (3-a) can be interpreted neither as a cosmetic preparation nor as a text marker, whereas *highlighter* in (3-b) may refer to any of the two[2]:

(3) a. *They still have the Stoke shirts in S and M and the Celtic **highlighter** abomination in XL.*
b. *Choose a **highlighter** that lets you be as precise as you like to be.*

Understanding that the phrase *highlighter abomination* refers to a football kit requires specific world knowledge, i.e. knowing at least that *Stoke* and *Celtic* are football teams. On the other hand, the meaning of *highlighter* in (3-b) is simply not fully determined given the available sentential context. To do away with a static notion of word senses, each usage of a word shall be considered as a unique modulation of that word's meaning, which can only be determined in context. Yet is there a natural language processing paradigm that allows for lexical meaning to be a priori underdetermined (Ludlow, 2014) and interpreted on the fly?

[1]Both sentences were found on the web as a result of querying for *highlighter* and *highlighter pen*. (2-a) is taken from `www.totalbeauty.com/content/gallery/best-highlighter`, while (2-b) is constructed with sentences from `www.jetpens.com/blog/the-best-highlighter-pens/pt/606`.

[2]Example sentence (3-a) is taken from r/LiverpoolFC (Section 3.1.2), while sentence (3-b) was found on a web page hosting reviews of text markers: `www.jetpens.com/blog/the-best-highlighter-pens/pt/606`.

Neural language models offer a way to produce token representations that are shaped dynamically by their sentential environment: for a given word type, each corresponding token representation is a learned function of the model's hidden layers, as activated by a sentence containing that word. In other terms, a language model assigns a different abstract representation to each of the four occurrences of *highlighter* seen above. Furthermore, the fact that language modelling involves sequential predictions rather than unstructured predictions (as e.g. Skip-gram (Mikolov et al., 2013a)) causes the internal states of the neural network to capture, in addition to semantic relatedness and functional similarity, also syntactic, compositional semantic, as well as information-structural properties of word distributions. Indeed, (pre-trained) neural language models have recently been shown to improve state-of-the-art performance in numerous natural language processing applications (Dai and Le, 2015; Peters et al., 2017, 2018; Radford et al., 2018; Howard and Ruder, 2018), including both sentence-level tasks such as natural language inference and paraphrasing, and token-level tasks such as named entity recognition and word sense disambiguation.

1.3 Thesis structure

The remainder of this thesis is structured as follows. The second chapter will provide the reader with an overview of linguistic theories of lexical semantic change as well as an excursus on different approaches to word representation learning. This will serve as a motivation for our methods, which we present in Chapter 4. Before, in the third chapter, we will present the raw language data, both historical and conversational, that our analyses are based on. In Chapter 5 we will describe how the proposed approach is evaluated as well as the results of our assessment. Then, the sixth chapter will showcase the types of analyses made possible by our approach, and it will discuss their successes and limitations. Finally, Chapter 7 concludes with a summary of our contributions and with considerations on the future potential of our method as well as on its usefulness to linguistic investigations and extrinsic applications.

Chapter 2

Background

The meaning of words and its diachronic change have been largely studied by historical linguists, lexicographers, lexical typologists, and more in general by scholars in the humanities and social sciences. The approach that these scholars have in common can be referred to as "close reading" (Moretti, 2013): it involves human reading[1] and manual analysis of texts. However standard and popular, this approach does rely on a crucial assumption: that a few important, paradigmatic texts, the *canon*, can be representative for the entirety of language production. This assumption was necessary until a few decades ago: indeed if the agent of linguistic analysis is human, there is no alternative to close reading. With the evolution of computer science, however, and the surge of the field of computational linguistics, new semi-automatic and automatic methods were introduced which could scale up this line of research, thereby moving from close to "distant reading" of texts. Distance is a "condition of knowledge" (Moretti, 2013) in the sense that it blurs away the peculiarities of a particular text and it allows the analyst to focus on units that are smaller or larger than the text. These methods have largely benefited from the digitisation of historical documents and from the alacrity of online language production. Both have been providing and are continuously contributing to a rapidly growing body of texts for distant reading. These corpora can span years, decades, centuries, and they provide support for quantitative analyses and testing of linguistic hypotheses, thereby allowing investigations into how the meaning of words changes over shorter or longer time spans as well as across speech communities. The combination of these two advances has given rise to an increasing number of studies on lexical semantic change which deal with its detection, characterisation, modelling, and generalisation (e.g., in the form of laws of semantic change (Dubossarsky et al., 2015; Xu and Kemp, 2015; Hamilton et al., 2016)) and which rely on formal, automatic, quantitative, and reproducible evaluation (Tahmasebi et al., 2018).

2.1 Lexical semantic change

Lexical semantic change can be studied from two main perspectives (Grondelaers et al., 2007). One is *onomasiological*, from function to form: onomasiological studies lay their focus on a referent, an object or an idea, and analyse the synchronically and diachronically varying ways of designating that referent. The other view is *semasiological*, from form to function: semasiological studies focus on a linguistic expression and investigate the synchronic and diachronic variation of the objects and ideas that are designated by that expression. Most of the latest computational approaches to semantic change modelling adopt a semasiological point of view. This is because most methods rely on abstract numerical representations of words or phrases that are obtained in a data-driven fashion, and it is not yet clear how to extract concept representations from data in the same bottom-up manner. Concepts are not spelled out in raw language data.

From the semasiological perspective, semantic change occurs when an existing form acquires or loses a particular meaning. In this sense, meaning change is strictly related to the evolution of the senses of a word form: its polysemy can increase or decrease as more or less referents are designated by the same form over time (Traugott, 2017). As an example, the word *virus* has made its first appearance in Late

[1] Intended as the common act of reading that literate humans perform on a daily basis.

Middle English texts with the meaning of snake *venom*. Subsequently it acquired the medical sense of a disease-related infective body substance. Only in the last century has it been used in its nowadays perhaps most prototypical sense of submicroscopic infective agent, and its most lately acquired meaning is that of a self-replicating malicious computer program. This type of change is also referred to as *semantic shift*.

Onomasiological studies often take a top-down approach and use resources such as ontologies (e.g. WordNet (Miller, 1995)) to fixate the concepts for which changing referential expressions will be tracked. Yet work on the emergence of concepts and semantic categories is at its very beginnings (Dubossarsky et al., 2015; Schmelzeisen and Staab, 2019). An exquisitely controversial example that can be studied onomasiologically is the ongoing change of terminology used to denote indigenous peoples of the Americas: *Indian*, *American Indian*, then the introduction of *Native American*, later only *Indigenous*, and sometimes *Amerind* or *Amerindian*. This type of change is often referred to as *lexical replacement*.

From both angles of view, in the first half of the 20th century linguists have devoted much of their theoretical work to categorising different types of semantic change (Bréal, 1899; Stern, 1931; Bloomfield, 1933). The resulting categorisations have inspired a number of more recent studies (Blank and Koch, 1999; Geeraerts et al., 1997; Traugott and Dasher, 2001) and are described in modern textbooks on language change (e.g. Hock and Joseph, 2009; Campbell, 2013).

The main types of change—of which e.g. Traugott (2017) offers historical examples—are:

- *broadening* (or *generalisation*): the extension of the range of concepts designated by a term,
- *narrowing* (or *specialisation*): the contraction of the range of concepts designated by a term,
- *metaphorisation*: the conceptualisation of one referent in terms of another, guided by analogical reasoning and implying an unspoken simile,
- *metonymisation*: a meaning transfer from one word to another, guided by spatial, temporal or causal contiguity between the two referents,
- *amelioration*: the acquisition of or shift towards a positive connotation,
- *pejoration*: the acquisition of or shift towards a negative connotation.

As it is a result of the conventionalisation of interactional strategies within groups of speakers, semantic change is almost never abrupt; it rather involves a process of *polysemisation*. In other words, a shift from a word sense A to a new sense B never occurs directly—$[A] \rightarrow [B]$—but rather through an intermediate polysemous stage, such as $[A] \rightarrow [A, B] \rightarrow [B]$, or $[A] \rightarrow [A, b] \rightarrow [a, B] \rightarrow [B]$, where capitalisation is used to denote the dominant word sense (Kutuzov et al., 2018).

2.2 Word representations

A variety of methods has been proposed for the computational modelling of lexical semantics. To construct abstract representations of words, they rely on fundamental assumptions about language.

2.2.1 Distributional word representations

Distributional semantics approaches assume that the distributional hypothesis holds, i.e. that semantic similarity between words results in similarity of linguistic distributions (Harris, 1954). The idea is that if semantically related words occur in similar contexts (first-order co-occurrence), those contexts and their relative frequency can be used to induce semantic representations (Boleda, 2019). Co-occurrences can be modelled with count-based methods (Turney and Pantel, 2010; Baroni and Lenci, 2010) as well as with predictive neural models (Turian et al., 2010; Collobert and Weston, 2008; Collobert et al., 2011; Mikolov et al., 2013a; Pennington et al., 2014) which have nowadays largely gained ground thanks to their good performance (Baroni et al., 2014; Levy and Goldberg, 2014) and to three important properties of the representations they output (Boleda, 2019): (i) they are learnt unsupervisedly from raw

natural language data, (ii) their *multi-dimensionality* captures multiple nuanced—though not necessarily interpretable (Boleda and Erk, 2015)—aspects of meaning, and (iii) their continuous nature reflects gradedness in semantic phenomena such as word similarity, synonymy, lexical priming, and selectional preferences.

In distributional semantics models, the unit of representation is the orthographic form: i.e. to one word form corresponds one distributional representation. However powerful, this *type-based* approach requires explicit composition rules (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Socher et al., 2012; Mikolov et al., 2013b) in order to express meaning *in context*, thus it is intrinsically inapt for the modelling of polysemous words and of polysemysation processes. It is not feasible to accurately aggregate all senses of a word into a single representation when the distributional properties of two distinct senses do not overlap (cfr. Section 1.2). Another crucial downside of distributional word representations is that they discard sequential information tout court and only rely on collocations; they encode similarity of linguistic environments by modelling texts as unordered collections of words. Moreover, although directly deployable as word features in downstream tasks, distributional representations can only be interpreted in second-order terms, i.e. via examination of their nearest neighbouring words in a multi-dimensional semantic space.

2.2.2 Contextualised word representations

Distributional word representation learning hence produces static, context-independent word features. In contrast, high quality word representations should ideally also model how a word's collocational properties vary across different contexts. The first well known attempt to address this limitation is a clustering-based disambiguation algorithm for word usage vectors (Schütze, 1998). In this method, the occurrences of a given word are grouped together based on second-order co-occurrence, i.e. instead of constructing a context-sensitive representation for a word based on the words that directly occur with it, context-sensitivity is achieved using the terms that these words in turn co-occur with in the training corpus (Schütze, 1998). A few years after this pioneering work, more researchers have begun to focus on word sense induction and word sense representations (McCarthy et al., 2004; Reisinger and Mooney, 2010; Neelakantan et al., 2014) within the field of distributional semantics. Others have proposed to learn multiple vectors for the same word type by relying on the word's selectional preferences for its argument positions (Erk and Padó, 2008) and to directly learn usage-specific representations based on the set of exemplary contexts wherein the target word occurs (Erk and Padó, 2010). The latest, deep learning oriented approaches to learning context-dependent word features embed the representation learning task into neural machine translation (CoVe; McCann et al., 2017) or into language modelling (Dai and Le, 2015, ULMFiT; Howard and Ruder, 2018, ELMo; Peters et al., 2018, GPT; Radford et al., 2018, 2019, BERT; Devlin et al., 2019). This paradigm shift has the potential of modelling not just collocational but also collostructural characteristics of word use (Stefanowitsch and Gries, 2003; Gries and Stefanowitsch, 2004; Goldberg et al., 2004): lexemes are analysed in interaction with the grammatical structures wherein they are embedded.

As an example, Peters et al. (2018) propose a technique to obtain deep contextualised word Embeddings from Language Models (ELMo), which can be easily integrated into a variety of task-specific NLP architectures. ELMo is a stack of bidirectional LSTMs first trained as a language model and then augmented with task-specific layers. This type of modular architecture allows sequential information to be processed by the model left-to-right and right-to-left, and it encourages the model to hierarchically distribute across layers the detection and processing of different lexical and sentential features. Indeed the transferability of ELMo layers has been investigated in recent work (Liu et al., 2019) and higher ELMo layers have been shown to be mostly tailored to higher level linguistic properties (such as those regulated by long-distance dependencies) and to context-dependent aspects of word meaning (Peters et al., 2018), whereas lower layers appear to encode simple word and sentence features (low-layer hidden states represent some aspects of syntax and they can be successfully used e.g. for POS tagging).

As it is a language model, ELMo's goal is to predict the next most likely word given a sequence of tokens. Although word representation learning is not treated as an explicit learning objective, ELMo's

layered structure provides a handle on the LSTM activations that correspond to a specific input word, so that the representation of each input token is a function of the entire sentence that contains it. Unlike previous approaches (Peters et al., 2017; McCann et al., 2017), the learned representations are *deep*—in the sense that they are a learnable[2] linear combination of all the internal layers of the language model—and they are *contextualised*, as it has been verified in zero-shot word sense disambiguation settings, where raw ELMo word features yield results that are on par with state-of-the-art WSD models (Raganato et al., 2017; Peters et al., 2018).

On the other hand, following the trend that developed about a decade ago in computer vision, NLP researchers have also begun to rely on large neural architectures trained using abundant language data and without any human supervision. Once trained, these neural networks can be deployed with little or no fine-tuning to token-level tasks (e.g. named entity recognition), sentence-level tasks (e.g. sentiment analysis) as well as tasks that require inter-sentential reasoning (e.g. paraphrase detection and natural language inference). This pretraining approach was followed by researchers at the University of Ulm (Howard and Ruder, 2018), OpenAI (Radford et al., 2018, 2019), and Google (Devlin et al., 2019), who proposed similar attention-based language modelling architectures. A crucial characteristic of attention-based models (Vaswani et al., 2017), which differentiates them from recurrent neural models and makes them suitable for parallelised computation, is the fact that they do not process sentences sequentially. Instead, stacked attention layers build dynamic representations for a target word in terms of the relation of that word to all other input tokens (and sometimes to the target token itself). In this way, the model's access to long-distance cues is not mediated by the processing of intervening sentential material, as it is in the case of recurrent models. Another point of contrast with recurrent contextualising models is that attention-based architectures tend to lack a single most transferable layer. The best performing layer varies across tasks—it is usually towards the middle—and a linear combination of layers typically outperforms any individual layer (Liu et al., 2019).

In the current work, we describe and deploy Google's BERT (Bidirectional Encoder Representations from Transformers) as it was shown to achieve the best performance on various tasks as well as on pure language modelling (Devlin et al., 2019; Liu et al., 2019). The main novelty of BERT is a *masked* learning objective that allows the representation to "fuse" left and right context (Devlin et al., 2019): the new masked language model, inspired by Taylor's Cloze test (Taylor, 1953), masks multiple input tokens in a sentence and its objective is to predict the correct words for the masked slots, based only on sentential context. Besides this token-level task, BERT is also trained on a binary next sentence prediction task, which forces the model to capture relationships between sentences.[3] This is particularly useful for sentence-pair tasks such as paraphrasing and natural language inference.

The intuition behind BERT's double training regime is that filling randomly positioned slots and recognising connected sentences requires awareness of the meaning carried by lexical items, the meaning carried by grammatical and rhetorical structures, socio-cultural meaning, as well as simple *situation models* (Fries, 1963; Kintsch, 1988). As a consequence, BERT can also be used effectively as a word representation learner in order to obtain rich and truly contextualised features.

2.3 Semantic change modelling

2.3.1 Type-based approaches

The three key properties outlined in Section 2.2.1 make distributional semantics a viable framework for the automatic analysis of semantic change: multi-dimensionality of word representations allows for the modelling of nuanced semantic shift, their gradedness reflects well the continuous nature of change, and the representation learning process does not require any data annotation (Boleda, 2019). The standard distributional approach to semantic change modelling is to separately train distributional models on the time bins that constitute a corpus (Gulordava and Baroni, 2011) and to measure distance between representations obtained for the same word with diachronically trained models. Representational coherence between features obtained for adjacent periods can be guaranteed by incremental training procedures

[2]The linear parameters are optimised with respect to an extrinsic task.

[3]The final pre-trained model achieves ca. 98% accuracy on this task.

(Kim et al., 2014) as well as by post hoc alignment of semantic spaces (Hamilton et al., 2016). Alternatively, some methods learn relations from word usages to time periods directly (Bamler and Mandt, 2017; Rosenfeld and Erk, 2018; Rudolph and Blei, 2018). What all these approaches have in common is the shared assumption that meaning change results in change of linguistic distribution, measured as first-order co-occurrence.

Proceeding chronologically, one of the first approaches to the automatic quantification of change based on diachronic corpora is the one proposed by Michel et al. (2011). It is based on the idea that significant growth in the relative frequency of a word can be an indicator of semantic shift. In the case of generalisation and narrowing, for example, the acquisition or loss of a word sense usually correspond to an increase or decrease in the raw frequency of the word in the corpus. Frequency change is, however, a very loose approximation of semantic change and it yields a large amount of false positives, i.e. words whose frequency in the corpus has changed but whose meaning has remained constant. Gulordava and Baroni (2011) move from frequency-based measures of change to a distributional model of lexical semantics. They build a co-occurrence matrix using a fixed context size and fill its values using Local Mutual Information (Evert, 2008). Semantic shift is characterised as variation in distributional similarity, expressed as the cosine distance between feature vectors obtained for a word in two adjacent time periods. The vocabulary (hence the vector dimensionality) is fixed across time periods, so that the resulting word representations lie in the same vector space. Low self-similarity across decades indicates semantic change.

To extend this methodology to more than two periods, Kim et al. (2014) introduce an incremental training procedure for diachronic word representations. Given a corpus of texts divided by period of production, a Skip-gram model receives as input the texts from period *t* and outputs epoch-specific word embeddings. The obtained vectors are used to initialise the Skip-gram embedding matrix at $t + 1$. To identify the specific periods during which change has occurred, this method relies on the geometric self-similarity of the representations obtained for a word over time: $cos\left(v_w^t, v_w^{t+1}\right)$. An important limitation of this technique is that it is biased with respect to a word's frequency of occurrence: if usages of a word decrease dramatically starting from $t = x$ (as in reductive semantic change or sense loss), word vectors for $t > x$ will remain virtually the same and semantic change will remain undetected. The authors suggest combining cosine distance and frequency to define a new more robust metric.

If, as observed by Kim et al. (2014), a word's different frequencies of occurrence across time periods can cause distributional models to fail at detecting semantic changes (e.g. complete loss of a word sense), surges in frequency are problematic too: they do not always indicate actual change and can occur e.g. as a result of real-world events. Therefore, to complement insights from distributional similarity and word frequency analyses, Kulkarni et al. (2015) also investigate whether a word's part-of-speech changes over time. Correspondingly, they construct three different types of time series and propose a statistically sound change point detection algorithm.

- Frequency-based time series are built using the log probability of occurrence of a word for each epoch-specific snapshot of the corpus. This method is sensitive to bias in domain and genre distributions, and to sudden or unpredictable popularity shifts of specific entities and events.

- Syntactic time-series are constructed measuring the Jensen-Shannon divergence between POS distributions across successive snapshots. The intuition is that when a word acquires a new sense, its syntactic environment in the corpus will vary (e.g. *to download* vs. *a download*) to reflect the acquisition of new syntactic functionalities.

- Distributional time series are constructed by computing the cosine distance between vectors of the same word obtained in different epochs. Skip-gram is used to learn word representations: (i) for each time period the model is initialised with random embeddings, then (ii) the model is trained independently for each time period, and finally (iii) the vector spaces are aligned to the final snapshot by learning a linear transformation mapping every word from an embedding space to the successive one.

Given any of these time series, a Mean Shift model (Taylor, 2000) is used to determine if a word has changed significantly and, if so, what the exact change point is. Kulkarni et al. (2015) argue that even if change in word meaning happens gradually, a time period can be identified where the new usage takes over (a *tipping point*).

Unlike the incremental training approach of Kim et al. (2014) but similarly to that of Kulkarni et al. (2015), Hamilton et al. (2016) propose a method to force alignment of diachronic embeddings to the same coordinate axes. They use orthogonal Procrustes to learn an optimal rotational alignment between the word embedding matrix $\mathbf{W}^t$ and the matrix $\mathbf{W}^{t+1}$ obtained in consecutive time period. As a measure of semantic shift, they propose second-order similarity between word representations. This is obtained by computing pairwise similarity over time with respect to a selection of prototypical lexemes: for a word w, an ordered vector of cosine similarities $\boldsymbol{sim}(w, *, t)$ is computed for each time period and compared to $\boldsymbol{sim}(w, *, t')$ using Spearman rank-correlation coefficient.

2.3.2 Sense-based approaches

There exist extensions of distributional representations that use senses as their unit of meaning (Chen et al., 2014; Neelakantan et al., 2014; Wu and Giles, 2015; Liu et al., 2015). Nonetheless multi-sense embeddings do not consistently improve over type-based ones (Li and Jurafsky, 2015) and they have not been used to track the evolution of senses over time.

Alternative unsupervised approaches have been proposed that do not directly rely on multi-sense embeddings but which still allow for the modelling of polysemous words. A prominent example of this line of work is the noun sense identification pipeline proposed by Mitra et al. (2014, 2015). They use Google Books (Michel et al., 2011) to produce epoch-specific distributional thesauri (Rychlỳ and Kilgarriff, 2007): the dependency-parsed contexts of each word and the frequencies of the syntactically annotated contexts are used to calculate the lexicographers mutual information (Kilgarriff et al., 2004) between a word and its contextual syntactic features. Using a co-occurrence-based graph clustering framework (Biemann, 2006), the top 1000 contextual features of each time period are grouped together, so that each cluster hypothetically corresponds to a word sense. If a word undergoes sense change, this can be detected by comparing sense clusters obtained from two different time periods. Such comparisons can reveal the birth of a new sense, the death of an existing one, as well as the split of a single sense into multiple senses and the formation of a new sense due to the combination of two older senses. The *stability* of a sense—whether it is uninterruptedly detected across time spans—, its *age*—the number of time periods wherein it has occurred—and the *location* of a change—the time period where the change is first detected—can also be measured. An important property of this approach is that it starts taking collostructural word features into account. However, in order to do so, it requires the texts to be syntactically parsed. Moreover, several threshold values and heuristics need to be introduced in order for sense change to be reliably detected; most notably the set of candidate words for semantic change detection is filtered down to include only nouns.

With the potential of being methodologically and statistically more principled, a number of Bayesian models of meaning change have been also developed (Wijaya and Yeniterzi, 2011; Lau et al., 2012, 2014; Cook et al., 2014). The latest and so far most successful one is SCAN, a Bayesian model of sense change (Frermann and Lapata, 2016). Conceptually similar to (Lau et al., 2012) and inspired by dynamic topic models (Blei and Lafferty, 2006), SCAN models the meaning of a word as a set of senses which change their relative prevalence over time, and it assumes that temporally adjacent representations are co-dependent in order to guarantee smoothness of semantic change. Each target word is modelled separately, using a collection of fixed-size context windows annotated with their period of origin.[4] For each time period approximate inference results in a distinct word representation, which is defined as (i) a multinomial distribution over K word senses, (ii) a $|V|$-dimensional distribution over the vocabulary for each word sense $k \in [1, \ldots, K]$, and (iii) a precision parameter which regulates the variability of temporally adjacent representations (Frermann and Lapata, 2016). The temporal representations induced by SCAN can be successfully deployed for the detection of meaning change between two time periods

[4]The length of the time intervals (temporal granularity) can be set as a hyperparameter.

as well as for the identification of a text's epoch of origin.

Although it can model polysemy as the co-existence of multiple latent word senses and polysemisation as gradual change in the relative prevalence of senses, SCAN presents two important limitations. First, it relies on the manual setting of the number of word senses K, which is the same for all words in (Frermann and Lapata, 2016) and remains constant across time as a result of temporal co-dependence. Second, it treats context as a bag of words—whose size must be therefore also chosen as a hyperparameter. Selecting K in this manner is equivalent to assuming (i) that every word possesses multiple senses, (ii) that every word form corresponds to the same amount of senses, and (iii) that the number of senses can be preemptively guessed by the modeller. Even if the second assumption were to be dropped (which is indeed possible with SCAN), an anticipated decision with regard to the number of senses that compose a word's meaning is still required. The latter aspect can be particularly problematic for corpora with community specific language use. As an example, the word *highlighter* is used in an online community of Liverpool FC fans to refer to a particular fluorescent-yellow football jersey (cfr. Section 1.2); its prototypical, community-agnostic interpretation of broad pen used for marking documents is absent—not to mention the nowadays perhaps even predominant cosmesis-related sense. The second drawback, which is shared with traditional type-based models, is that SCAN solely relies on the distributional hypothesis. That is, it assumes that the meaning of a word can be exhaustively modelled via the word's relatedness to co-occurring lexemes as well as its similarity to lexemes that exhibit similar corpus distributions. The sentential context of a word of interest is simply expressed as an unordered collection of tokens occurring within a limited, fixed distance from the word. As we have discussed in Sections 2.2.1 and 2.2.2, dispensing with the inherently sequential structure of sentences results in the inability to capture compositionality, long distance syntactic and semantic relations, as well as more global properties such as topic and information structure.

2.3.3 Towards a usage-based approach

To discard these assumptions, we adopt a theory of lexical semantics that deems word meaning as inherently underdetermined and contingently modulated in situated language use. Every usage of a word must undergo an interpretation—at least by the speaker, hopefully by the reader—that is necessarily shaped by the word's context of occurrence. In other terms, if one representation for each word form is obviously not sufficient for the accurate modelling of lexical semantic features, defining a fixed number of underlying senses is only a refinement of the first method and it still amounts to considering polysemy as a discrete and static phenomenon (Section 1.2). To address this limitation, we rely on a neural language model and obtain contextualised representations that are uniquely defined by a word form together with its *entire* sentential context. In the proposed approach there is no need to define in advance a number of senses—salient types of word usages emerge from the data. Furthermore, usage types are represented directly via the usages that define them. Second order descriptions are unnecessary as each word interpretation is defined in terms of the sentence wherein the word occurs. Similarly, each emerging usage cluster can be represented abstractly by the respective cluster centre as well as by the sentence that generated the closest contextualised word vector to that centroid. This approach results in a much more human-friendly, interpretable characterisation of word meaning. Another advantage of using language models to obtain word representations is that they encode more than shallow topic relatedness and explicit distributional similarity. By factoring in sequentiality, these models encode long distance dependencies as well as global properties of sentences.

Chapter 3

Natural language data

Historical data sets are of prime relevance for the analysis of language change. Available diachronic corpora can be broadly categorised into those that span long periods of time (e.g. multiple decades, centuries) and those that cover shorter periods (e.g. months, a few years). The longitudinal extension of a corpus contributes to determine the types of semantic change that can be detected and analysed. Linguistically motivated semantic shifts tend to be found, for example, in long-term resources, whereas short-term corpora are useful for analysing socio-cultural semantic drifts (Kutuzov et al., 2018).

The most prominent long-term resource is probably the Google Books Ngrams corpus (Michel et al., 2011). It covers 5 centuries, from 1520 to 2008, and it has been used in numerous studies to detect differences in word meaning and connotation across arbitrarily wide time spans (e.g. Gulordava and Baroni, 2011; Mitra et al., 2014). A disadvantage of this corpus is that Google Books texts are distributed in n-grams and the rarest n-grams have been discarded. This can be problematic for methods, such as ours, that require the processing of entire sentences rather than of a restricted context window surrounding the word of interest. Another well established resource is the Corpus of Historical American English (Davies, 2012), which spans two centuries, includes full texts from four different genres, and is genre-balanced decade by decade.

Examples of corpora that span shorter time periods include the Corpus of Contemporary American English (Davies, 2010), containing genre-balanced texts from 1990 to 2017, and the New York Times Corpus (Sandhaus, 2008), with news articles from 1990 to 2016. As the temporal extension of the corpora decreases, the granularity of the time spans typically increases, allowing to explore faster-paced meaning negotiation dynamics (Clark, 1996; Hasan, 2009). An increasing number of Computer-Mediated Communication data sets has indeed made its appearance in the field. E.g. Kulkarni et al. (2015) use Amazon Movie reviews with a granularity of 1 year as well as Twitter data with a granularity of 1 month, and Del Tredici et al. (2019) deploy a dataset of conversations occurred over a period of 8 years on an online forum.

3.1 Diachronic data sets

3.1.1 Historical corpora

The Brigham Young University has made available two unique resources for the study of historical and contemporary American English. Their diachronic nature and large size make these corpora particularly apt for the investigation of changes at all linguistic levels: lexical, morphological, syntactic, semantic, and discursive. We will present COHA and COCA in the following sections.

COHA

The Corpus of Historical American English (COHA; Davies, 2012) is one of the largest resources for the study of variation and change in American English. It consists of approximately 400 million words and it covers two centuries of language use: from 1810 to 2009. Texts are canonically divided into decades but they are annotated on a year by year basis. After the 1880s, approximately 2 million words are available for each year.

The corpus is not only structured longitudinally. Texts of four different genres were collected: *fiction*, popular magazines (*magazines*), *newspapers*, and *non-fiction*. They were assembled from a variety of sources, including archives such as Project Gutenberg[1] as well as scanned and PDF documents[2], movie and play scripts. The compilers balanced the corpus decade by decade, so that the relative frequency of the four genres is approximately constant across time bins.[3] The absolute genre distribution, on the other hand, is not uniform as *fiction* accounts for ca. 50% of the texts available in each decade.

COCA

The Corpus of Contemporary American English (COCA; Davies, 2010) is an even larger resource for the study of contemporary language use, its variation and change. Texts are organised on a yearly basis and they were collected from 1990 to 2017, for a total of 560 million words. The span of this corpus is shorter compared to COHA's while the granularity is the same, yet the amount of data available for each time bin is at least 10 times larger: for each year approximately 20 million words are available. COCA's more than 160,000 texts are uniformly balanced by genre—though not on a year by year basis—and three genres overlap with those found in COHA: (i) *fiction*, consisting of short stories, children's magazines, first chapters of books, movie and play scripts, (ii) *magazines*, including a selection of almost 100 magazines from a variety of domains—from *Good Housekeeping* to *Fortune*, and (iii) *newspapers*, including different sections of 10 US newspapers. Non-fiction is replaced by (iv) *academic journals* and (v) transcribed conversations are included as an additional genre (*spoken*).

3.1.2 Conversational corpora

r/LiverpoolFC

The r/LiverpoolFC corpus (Del Tredici et al., 2019) was created as a resource for short-term meaning shift analysis. It covers 8 years of language use in an online community of speakers hosted by the Reddit forum platform from 2011 to 2017. The subreddit[4] under consideration is r/LiverpoolFC, one that gathers fans of the English football team. This dataset was compiled with the idea of providing researchers (i) with texts organised according to sufficiently high temporal granularity, so that abrupt shifts can be detected, and (ii) with the language use of a specific community, where non-standard word interpretations are more easily adopted. In addition, the social graph that connects r/LiverpoolFC redditors exhibits high density, a characteristic that makes it a better environment for the fostering of linguistic innovations (Del Tredici and Fernández, 2018). The size of the corpus is not negligible—it consists of 40 million words—and each utterance is annotated with a timestamp, enabling analyses at a custom level of granularity. One disadvantage is that texts are non-uniformly distributed across time, as a result of the increasing popularity of the r/Liverpool subreddit. This imbalance can make it hard to study changes that occur in the first months covered by the data set, which only contain a few user posts.

Reddit 2013

As well as a community-specific collection of online discussions we use a large community-independent sample of users posts. This additional corpus can help models specialise on language use *in conversations* regardless of the topic of discussion. The Reddit 2013 data set consists of timestamped posts crawled from multiple subreddits, monthly, in 2013.[5]

3.2 Evaluation data sets

We have so far described the diachronic corpora used as training data for our semantic change analysis. These data sets provide, for every text, the exact date and time or at least the interval indicating when the text was produced. What they lack is an annotation that specifies which words have actually undergone

[1] www.gutenberg.org

[2] Scanned with Optical Character Recognition from printed sources, these documents also went through a post-processing phase. This clean up process was not sufficient, however, to prevent some texts from being quite unnatural to read.

[3] With the exception of the first 5 decades, which do not include newspapers.

[4] A *subreddit* is an online forum hosted on www.reddit.com where users discuss a particular topic.

[5] The data was downloaded from http://files.pushshift.io/reddit/comments/.

semantic shift during the intervals covered. Rather than relying on synthetically generated sets (as in e.g. Cook and Stevenson, 2010; Kulkarni et al., 2015; Rosenfeld and Erk, 2018) we use human-annotated lists of semantically shifted words, ranked by the degree of their shift.

For our analysis of lexical semantic change in COHA, we use as a reference the human judgements collected by Gulordava and Baroni (2011). From a set of 10,000 randomly selected mid-frequency words, they chose 100 words from different frequency ranges. This shorter list was then ranked by human raters according to their intuitions about semantic change from the 1960s to the 2000s, using a 4-point scale (0: no change; 1: almost no change; 2: somewhat change; 3: changed significantly). Annotations were averaged to produce a continuous shift score. The inter-annotator agreement, measured as the average pair-wise Pearson correlation, was 0.51.

On the other hand, an important benefit of the r/LiverpoolFC corpus (Section 3.1.2) is that it comes with a list of words annotated by Reddit users of the r/LiverpoolFC community itself.[6] The evaluation dataset consists of 100 word forms; 34 of these were identified as shift candidates by the authors[7] while the rest are confounders: 33 word forms that underwent a frequency increase from 2011-2013 to 2017, and 33 with constant frequency. A total of 26 Reddit users were shown the 100 words and, for each word, they provided a binary annotation indicating whether change had occurred. This annotation process yielded an average of 8.8 judgements per word, which were then aggregated into a *semantic shift index* by averaging. Shift index values range from 0 to 1 and a value higher than 0.5 indicates that the majority of raters who expressed their vote consider semantic shift to have occurred for the word under consideration. The inter-annotator agreement, measured as Krippendorff's alpha, was 0.58.

[6] Corpus and annotated data set are available at `github.com/marcodel13/Short-term-meaning-shift`.

[7] *"Semantic shift is defined here as a change in the ontological type that a word denotes, which takes place when the word starts to be used to denote an entity which is different from the one originally denoted and the new use spreads among the members of a community"* (Del Tredici et al., 2019).

Chapter 4

Methods

In Chapter 2 we have presented lexical semantic change as a linguistic phenomenon and we have discussed three types of approaches to the analysis of lexical semantic change: one is type-based and it does not take word polysemy into account, the second focuses on word senses but it assumes that the number of underlying word senses can be established a priori, and ours, the third, is based on unique contextualised word usages. A usage-based approach allows us to model the meaning of words as underdetermined, and to determine the unspecified aspects of word meaning on the fly, drawing information from the contingent sentential context of a word. In Chapter 3 we have presented the data sets on which our experiments rely, and thus we have clarified that the only type of supervision required by our procedure is the temporal annotation of the corpus of texts.

In this chapter, we finally present our method. We begin in Section 4.1 with a description of the deployed representation learning algorithm, the BERT language model (Devlin et al., 2019). In particular, we give an overview of the data generation process and of the model architecture, which are not explicitly discussed in (Devlin et al., 2019), and we describe two fine-tuning procedures: domain-adaptation and diachronic tuning. Domain-adaptive fine-tuning has the goal of adjusting the language model latitudinally, i.e. to the peculiarities of language use of a speech community or writing genre. On the other hand, diachronic fine-tuning is a training regime for longitudinal adaptation which produces period-specific language models.

Regardless of the tuning level (frozen, domain-adapted, or diachronically fine-tuned) BERT is then used as a language model to obtain contextualised word representations (or *usage representations*) for a list of words of interest. As a next step, all the usage representations collected for a given word form are aggregated into interpretable groups (*usage types*) via two clustering algorithms, K-Means and Gaussian mixture models. In Section 4.2, we provide a task-specific characterisation of both algorithms and present multiple ways of determining, in a data-driven fashion, the number of partitions they should yield.

The resulting clusters of usages, however, do not offer a diachronic view of word meaning. To include the temporal variable into our analysis, we propose a way of organising clustered word usages along the time axis, as described in Section 4.3. Lastly, although the resulting series of usage distributions are particularly apt for qualitative analysis via expressive visualisations, we propose metrics to quantify the distance between temporally contiguous usage distributions as an empirical measure of lexical semantic change (Section 4.4).

4.1 Language model

BERT is a multi-layer bidirectional Transformer encoder (Devlin et al., 2019) trained with two language modelling objectives: masked token prediction and next sentence prediction. The acronym BERT is often used to denote, in particular, a version of the language model that was trained on the BooksCorpus (800M words) (Zhu et al., 2015) and on English text passages extracted from Wikipedia (2,500M words). We refer to this pre-trained version as *frozen* BERT.

In the following sections we describe how to generate training data for BERT, its neural architecture,

as well as two tailored training regimes.

4.1.1 Data generation and processing

The BERT language model processes sentences in pairs, separated by a special `[SEP]` token. The first token of each sequence is the `[CLS]` symbol, whose hidden activation is used as a sentence representation for classification tasks—similarly to the last activation of a recurrent language model. Each input token is represented by the sum of three learned embeddings.

- *Word type embeddings*[1]: these are WordPiece embeddings (Wu et al., 2016) with a vocabulary of size $|V| = 30,000$. Words are segmented into character n-grams using a Wordpiece model which, given a training corpus and a desired vocabulary size $|V|$, selects $|V|$ types so that the size[2] of the tokenised corpus is minimal. A special symbol *##* is used at the beginning of non-initial word segments.
- *Positional embeddings* with a maximum supported sequence length of 512; i.e. $V = \{0, 1, ..., 512\}$
- *Segment embeddings*, with a vocabulary $V = \{A, B\}$, to designate whether an input token belongs to the first or to the second sentence of the training pair.

To train the model on the newly introduced masked language modelling objective (Section 2.2.2, Devlin et al., 2019) 15% of each sentence is masked according to the following procedure:

- 80% of the time, the `[MASK]` token is actually used
- 10% of the time, a random word replaces the masked token
- 10% of the time, the observed word is maintained.

BERT's second task, next sentence prediction (Section 2.2.2), is trivially generated from the training corpus: for each sample, when choosing the pair of sentences A and B, 50% of the time B is the sentence that actually follows A, and 50% of the time it is a random sentence from the corpus (Devlin et al., 2019).

The entire pre-processing procedure to generate training examples involves multiple steps. We follow the methodology used in the original BERT paper and repository[3], and treat our corpora as lists of documents: for conversational data sets, a document consists of a thread (a titled discussion within a subreddit), whereas the historical corpora are already conveniently distributed in single documents. Preprocessing is described in detail in Appendix A.

4.1.2 Model architecture

Each layer of BERT consists of a full Transformer block (Vaswani et al., 2017) which computes, given an input sequence of token representations $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, a continuous representation of the input $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$. A Transformer block is composed of two sub-layers:

(i) a multi-headed self-attention (or intra-attention) mechanism,

(ii) a fully connected feedforward layer consisting of two linear transformations and a non-linearity: $\mathrm{ReLU}(\mathbf{x}\mathbf{W_1} + \mathbf{b_1})\mathbf{W_2} + \mathbf{b_2}$.

Residual connections (He et al., 2016) and layer normalisation (Lei Ba et al., 2016) are applied to both (i) and (ii) such that the final output of each sub-layer is given by $\mathrm{LayerNorm}(\mathbf{x} + \text{sub-layer}(\mathbf{x}))$.

The attention mechanism consists of query, key, and value matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V}$, which are combined as follows to obtain an output matrix:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

[1] Note that this is referred to as *token embedding* in (Devlin et al., 2019). However, we find that naming misleading as there is only one such embedding for a given word form.

[2] The number of tokens.

[3] `github.com/google-research/bert`

where d_k is the dimensionality of the key vectors and $\frac{1}{d_k}$ is a scaling factor that differentiates the attention mechanism implemented in the Transformer, named Scaled Dot-Product attention, from standard Dot-Product attention. For multi-head self-attention, before the attention function is applied, the matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are linearly projected h times, using $3h$ learned affine transformations $\{\mathbf{W}_i^Q\}_1^h, \{\mathbf{W}_i^K\}_1^h, \{\mathbf{W}_i^V\}_1^h$. Then the projections are concatenated and their concatenation is again projected to a single output matrix:

$$\text{MultiHeadAttention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = [\mathbf{H}_1; \ldots; \mathbf{H}_h]\mathbf{W}^O$$
$$\text{where} \quad \mathbf{H}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right)$$

For classification tasks, a classification layer is added on top of the Transformer's final hidden layer and it is then connected to a softmax layer over the classes of interest. All the parameters of the Transformer blocks as well as, when available, the classification weights are fine-tuned jointly. Two versions of BERT were released: a smaller one, $\text{BERT}_{\texttt{BASE}}$, with 12 layers, 768 hidden dimensions, and a total of 110M parameters, as well as a larger one, $\text{BERT}_{\texttt{LARGE}}$, featuring 24 layers, 1024 hidden dimensions, and 340M parameters.

Devlin et al. (2019) give some suggestions with regard to the optimal strategy for the aggregation of neural activations across layers, based on results in a named entity recognition task. In our experiments we collect the activations of all of $\text{BERT}_{\texttt{BASE}}$'s layers and sum them dimension-wise. We do so for computational efficiency and because, in our preliminary analysis, neither selecting a subset of the layers nor using concatenation instead of addition produced notable differences in the relative geometric distance between word representations. Discrepancies solely emerge when only the lowest BERT layers are considered—in any case, such a selection would be difficult to motivate in light of the current understanding of stacked layer processing (e.g. Liu et al., 2019).

In the remainder of this thesis, we will use the acronym BERT to refer to $\text{BERT}_{\texttt{BASE}}$, and our BERT representations will be the 768-dimensional vectors obtained by summing $\text{BERT}_{\texttt{BASE}}$'s 12 layers.

4.1.3 Fine-tuning

BERT is a *pre-trained* language model. Its very large size together with the large amount and variety of training data are supposed to guarantee good performance on different tasks and different domains without excessive fine-tuning. In many applications, in fact, the frozen BERT performs on par with state-of-the-art models that have been specifically trained for the task at hand. Due to BERT's recency, however, it has not yet been studied in depth how to perform task-specific training optimally, neither is it clear when such a training is necessary.

Computer Vision pre-trained models, which have undoubtedly inspired the birth of pre-trained language models, have been object of a large number of studies which have tried to answer such questions. Should the model be fine-tuned for a specific task? Should it be adapted to new target domains? And if so, how should additional training be performed? E.g. should all the model weights be updated during fine-tuning or only the ones belonging to the last layers? Or should they perhaps all be updated but to a different degree (gradual unfreezing)? The answers have been perhaps not theoretically satisfying yet they have enabled researchers and professionals to use CV models in practice and with success.

In NLP, in the lack of theoretical indications and of sufficient empirical evidence, whether and how to give neural models some additional training has been treated as an engineering problem. In classification-based tasks, for example, an additional fully connected layer is added on top of BERT's Transformer modules and specifically trained, together with the Transformer blocks, for the task of interest. Sometimes, however, e.g. when the target domain significantly differs from the training domain or when it is a very specific one, BERT is first trained on the new domain using a language modelling (LM) objective. The motivation is very intuitive: while task-specific training of classification layers helps the model learn how to exploit the rich features learned during its massive pre-training, LM fine-tuning adapts the model to community-specific language use.

Following previous approaches to the training of word representation models for diachronic change modelling (e.g. Del Tredici et al., 2019; Han and Eisenstein, 2019) we propose two fine-tuning regimes:

one performs domain adaptation and the other is responsible for diachronic fine-tuning. **Domain-adaptive fine-tuning** is based on BERT's two standard training objectives: masked language modelling and next sentence prediction. We fine-tune BERT for n epochs on these two tasks and obtain as a result a single BERT model with updated weights. **Diachronic fine-tuning** has the same learning objective as domain adaption but it requires a corpus that is divisible into time bins, with custom granularity. For every time bin an epoch-specific language model is obtained. Each of these model is initialised with the weights of the model that chronologically precedes it. The first model can be initialised either simply as the pre-trained BERT or with domain-adapted weights. The latter procedure is similar to the incremental training proposed by Kim et al. (2014).

4.2 Clustering contextualised representations

Given that BERT can be deployed as a language model to obtain abstract usage representations, we need an algorithm that finds clusters in a set of d-dimensional contextualised word representations. First, we approach this task using a non-probabilistic technique, the K-Means algorithm (Lloyd, 1982). Then we exploit Gaussian mixture models, whose discrete latent variables define assignments of data points to specific components of the mixture—our usage types.

The data set is the usage matrix $\mathbf{U}_w = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, which consists of N observations of a random d-dimensional variable $\mathbf{x}$, i.e. the N contextualised representations obtained for the N occurrences of a word w in the diachronic corpus under scrutiny.[4] Our objective is to partition the data into K clusters (Figure 4.1a). Following the intuition provided by Bishop (2006), we can think of a cluster as a group of usage representations whose distances from each other are smaller compared with the distances to usages outside the cluster. To formalise this notion, let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ represent the centres of the clusters (or cluster centroids): $\boldsymbol{\mu}_k$ is a d-dimensional vector and can be thought of as the prototypical word use associated with cluster C_k. The procedure to select the value of K is described in Section 4.2.3.

4.2.1 K-Means

The goal of the K-Means algorithm is to find (i) the assignment of data points to clusters and (ii) the set of cluster centres which minimise the sum of the squared distances of each data point to its closest centroid $\boldsymbol{\mu}_k$, often referred to as *distortion* or *inertia*:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, ||\boldsymbol{x}_n - \boldsymbol{\mu}_k||^2$$

where r_{nk} is a binary assignment indicator variable such that $r_{nk} = 1$ if $\boldsymbol{x}_n \in C_k$ and $r_{nk} = 0$ if $\boldsymbol{x}_n \notin C_k$. To minimise this objective function, the K-Means algorithm follows the iterative procedure of the *Expectation-Maximization* algorithm (EM) until there is no further change in cluster assignments or until a maximum number of iterations is reached. Convergence is assured by definition—though it may result in a local minimum of J (MacQueen et al., 1967)—and it requires a different number of iterations depending on the initial position of the cluster centres.[5]

To alleviate the influence of different initialisation values, we run Expectation Maximization $I = 10$ times with I sets of initial centroid positions. The latter are chosen to be distant from each other according to the *k-means++* method (Arthur and Vassilvitskii, 2007), leading to an improvement over random initialisation. The final clustering is the one that yields the minimum *distortion* value across all runs. Finally, in our experiments, we standardise the data such that each of the variables has zero mean and unit standard deviation (Bishop, 2006) and we use Elkan's variation of the E-step to further accelerate EM (Elkan, 2003).

[4] In our case, the random variable $\mathbf{x}$ may be considered as a type representation of word w.

[5] For a stopping criterion, we rely on the default values of the following implementation: `scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`.

4.2.2 Gaussian mixture model

During each E-step of K-Means optimisation, every contextualised representation is assigned to a single cluster, the one with the nearest prototypical usage (*hard assignment*). This may be problematic for certain uses of a word that cannot be fully said to belong to one usage type or another (cfr. Section 1.2). In such cases, it might be better to adopt a probabilistic approach as a way of expressing uncertainty about the appropriate assignment. Indeed, with *soft assignments*, a single usage can be deemed to belong to multiple clusters, though with different cluster membership strength.

Given the data set of contextualised word representations $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, we use a Gaussian mixture model to obtain a latent variable for each observation, which will serve as a soft indicator of cluster membership. Let us therefore introduce a one-hot K-dimensional binary random variable $\mathbf{z}$ such that $z_k \in \{0, 1\}$ and $\sum_{k=1}^{K} z_k = 1$, and whose marginal distribution is specified in terms of the mixing coefficients $\pi_k \equiv p(z_k = 1)$, $\forall k \in [1, K]$. We may think of the mixing coefficient π_k as the prior probability for a word use to belong to usage cluster C_k (i.e. how predominant is, in general, usage type k?). This marginal is used together with a Gaussian conditional distribution of $\mathbf{x}$ to define the joint distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$. The marginal distribution of $\mathbf{x}$ is a Gaussian mixture:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

To measure cluster membership strength, we rely on the conditional probability $p(\mathbf{z}|\mathbf{x})$. If π_k can be viewed as the prior for cluster k, then $\gamma\left(z_k\right) \equiv p\left(z_k = 1|\mathbf{x}\right)$ shall be considered as the posterior probability of belonging to usage type k after having observed the occurrence of the target word in context:

$$\gamma\left(z_k\right) = \frac{p\left(z_k = 1\right) p\left(\mathbf{x}|z_k = 1\right)}{\sum_{j=1}^{K} p\left(z_j = 1\right) p\left(\mathbf{x}|z_j = 1\right)} = \frac{\pi_k \, \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \, \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}$$

The EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2007) is used to estimate the mixing coefficients π_k and the parameters of the Gaussian distributions $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$. EM does so by maximising the logarithm of the data likelihood:

$$\ln p(\mathbf{U}_w|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \, \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\}$$

Each iteration of EM is guaranteed to increase the log likelihood and the algorithm is executed until there is approximately no change in the log likelihood, in the mixture parameters, or until a maximum number of iterations is reached.[6] The number of iterations required to reach (approximate) convergence is typically higher than for the K-Means algorithm, and each iteration is more computationally expensive. We therefore follow the common practice to initialise the Gaussian mixture model with a run of the K-means algorithm.[7]

4.2.3 Selecting the number of clusters

We have so far given the number of clusters for granted. In this section, we will discuss how to select the value of K for K-Means as well as for the Gaussian mixture model. A notable difference between these two clustering methods is that the minimum number of clusters that we can obtain with K-Means and its K selection methods is 2, whereas a Gaussian mixture can also only rely on a single component.

[6] For a stopping criterion, we rely on the default values of the following implementation: `https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html`.

[7] The means $\boldsymbol{\mu}_k$ and covariance matrices $\boldsymbol{\Sigma}_k$ of the Gaussian distributions can be initialised, respectively, to the cluster centres and to the sample covariances of the clusters found by the K-Means algorithm. The mixing coefficients π_k can be set to equal the fractions of data points assigned to the respective clusters (Bishop, 2006).

Usage A	At the day's end, the *users*' association concept had been backed only by the western big three which sponsored the idea, Australia, New Zealand, Italy. Behind the *users*' association, which is at best only a provisional device, lies the possibility of a boycott of the canal. He pushed ahead with the American plan to establish an association of *users*.
Usage B	However, the Macintosh platform is still not as universally supported as are PCs, and this puts *users* at a disadvantage. Google provided search services for *users* of both Yahoo and AOL, putting its brand name in form of millions of computer users. But in that case the entire machine was immersed in inert nonconductive liquid—not exactly a practical setup for home *users*.
Usage C	It's worth noting that the antibiotics *users* were, on average, older and heavier, had stronger family histories of cancer and were more likely to use hormone-replacement therapy. Amphetamine *users* often become heavily dependent on the drug, which can produce the symptoms of schizophrenia. Pot prohibition gives sporadic *users* the stigma of criminal records and makes young people cynical about law in general.
Usage D	Great financial harm would soon follow for all *users* of steel... Mr. Ford has announced that he will ask all industrial *users* of his coal to install furnaces that will remove only the gas, leaving a fuel unimpaired for domestic purposes. Kodak vows to listen closely to its customers, and now sends manufacturing employees on road trips to meet with professional *users* of film such as Hollywood producers to find out their needs.
Usage E	No, the headaches are going to come when veteran *users* install one of these beauties and then try to lead their computer lives as if nothing had changed. Since e-mail *users* change addresses and internet providers often, they say, the registry can't be kept current. Though downloading Atlas was rough going (more than an hour on a 14.4 modem), patient *users* were treated to a program stuffed with new applications, part of Netscapes plan to outdazzle and outperform Microsoft.
Usage F	The system to suppress competition by 1) boycotts, 2) price cuts (against plants refusing to play ball), 3) identical bids to cement *users*, and 4) opposition to the building of new plants. Sensuality, in turn, has an almost murderous force. Always there are the *users* and the used. Slave caravans seem to march across the top of every page like an endless frieze. Most interesting and important consequences of the fact that railroads are private ways while motor highways are open to the public is the contrast between the methods by which the *users* in each case bear the annual cost of the way.

Table 4.1: Usages of the word *users* in their context of occurrence (COHA). Each usage is among the five nearest observations to the respective cluster centre. Usage type clusters are obtained with K-Means clustering and the frozen BERT.

K-Means: silhouette and variance ratio

Although it may seem straightforward to use distortion as a criterion (Section 4.2.1), distortion is not a normalised metric, hence it will always yield the highest allowed K. We therefore rely on two metrics that are commonly employed to evaluate the quality of hard assignments: the silhouette score, and the variance ratio criterion. We choose these metrics as they do not rely on a known set of labels, which of course we lack for the current task.

The *silhouette coefficient* of an observation $\mathbf{x}$ is a measure of the quality of its assignment to cluster C_k. It is defined in terms of $a(\mathbf{x})$, the average distance between $\mathbf{x}$ and all other points in C_k, and $b(\mathbf{x})$, the average distance between $\mathbf{x}$ and all the data points in $\mathbf{x}$'s next nearest cluster:

$$s(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{\max\{a(\mathbf{x}), b(\mathbf{x})\}}$$

The silhouette coefficient is a value between 1 and -1 and it quantifies how close an observation is to data within the correct cluster and how far it is from data in the closest neighbouring cluster. The overall *silhouette score* of a clustering is the average silhouette coefficient over all clustered samples. It favours intra-cluster coherence and inter-cluster dissimilarity.

On the other hand, the *variance ratio criterion* (or *Calinski-Harabasz score*) for k clusters and N observations is defined in terms of the within-cluster dispersion matrix $W(k)$ and the between-cluster dispersion matrix $B(k)$:

$$\begin{aligned}
\text{VRC(k)} &= \frac{Tr(B_k)}{Tr(W_k)} \, \frac{N-k}{k-1} \\
B_k &= \sum_{i=1}^{k} |C_i| \; (\boldsymbol{\mu}_i - \mu(\mathbf{U}_w)) \, (\boldsymbol{\mu}_i - \mu(\mathbf{U}_w))^T \\
W_k &= \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \boldsymbol{\mu}_i) \, (\mathbf{x} - \boldsymbol{\mu}_i)^T
\end{aligned}$$

Gaussian mixture model: AIC and BIC

To choose the optimal number of mixture components, we can use criteria for model selection such as the Akaike Information Criterion and the Bayesian Information Criterion. Both methods are meaningless in

isolation: given a set of candidate models, they find which model provides the best approximation of the data. This involves determining which model guarantees the minimal loss of information with respect to the true data distribution.

Given a set of P model parameters $\boldsymbol{\theta}$ and the model likelihood $\mathcal{L}(\hat{\boldsymbol{\theta}})$ under the maximum likelihood estimate of $\boldsymbol{\theta}$, the Akaike Information Criterion (AIC; Akaike, 1998) is given by:

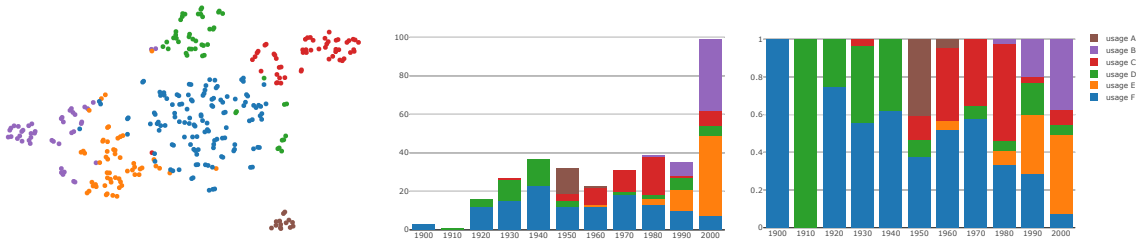
$$\text{AIC} = -2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}) + 2P$$

The subtrahend $-2 \log \mathcal{L}(\hat{\boldsymbol{\theta}})$ provides a measure of the model's goodness of fit, whereas the minuend $2P$ is a penalty for model complexity.

The Bayesian Information Criterion (BIC; Schwarz et al., 1978) is a slight modification of AIC which imposes a stronger penalty for the number of estimated parameters:

$$\text{BIC} = -2 \log \mathcal{L}(\hat{\theta}) + P \log N$$

The lower the AIC or BIC value, the better the trade-off between model fitness and model complexity. However, as it features a stricter regularisation term, BIC always selects models that are smaller or equal in size compared to AIC. It follows that, in general, AIC is better in situations when a false negative finding is considered more misleading than a false positive—e.g. when a single cluster contains word usages of two different types. The opposite holds for BIC: it is more adequate when false positives are deemed more problematic than false negatives—e.g. when two separately detected usage types should rather be merged.



(a) (b) (c)

Figure 4.1: T-SNE visualisation of the contextualised representations collected in COHA for the word *users* with the frozen BERT, coloured according to the usage type assigned to them by a K-Means clustering (a); the resulting diachronic usage cluster frequency (b) and probability distributions (c).

4.3 Usage type distributions

In the previous two sections we have described the language model used to obtain contextualised word representations and the clustering algorithms used to partition representations into an automatically determined number of usage types (Figure 4.1a). We now present a way of organising word usages along the temporal axis which is data-driven, generalisable to any collection of words, and which nevertheless yields lexeme-specific characterisations (Figures 4.1b, 4.1c). Each word of interest is modelled separately, hence it can possess a different number of usage types which are induced directly from language use and characterised by real sentences (Table 4.1).

We begin by collecting all the usages of a word of interest w that can be found in the corpus under scrutiny. For each usage of the target word in a sentence $s = \left(v_0, \ldots, w, \ldots, v_{|s|}\right)$, we store the corresponding contextualised representation $\text{BERT}(w|s) \in R^d$ output by the language model. The resulting set of N word usages is then represented as a usage matrix $\mathbf{U}_w \in R^{N \times d}$. Usage matrices can also be specific to the usages that occur in a certain time period $t \in [1, T]$; in that case, they are denoted as $\mathbf{U}_w^t$, such that $[\mathbf{U}_w^1; \mathbf{U}_w^2, \ldots, \mathbf{U}_w^T] \equiv \mathbf{U}_w$.

A straightforward approach to determining which types of usages occurred in an interval is to independently cluster the contextualised representations obtained in that interval. This approach results in T sets

of partitions, one for each time-specific usage matrix $\mathbf{U}_w^t$. We refer to these partitions as $C_1^t, \dots C_{K_w^t}^t$, with centroids $\boldsymbol{\mu}_1^t, \dots, \boldsymbol{\mu}_{K_w^t}^t$ (where K_w^t is the number of partitions obtained by clustering $\mathbf{U}_w^t$ according to a K selection metric, as discussed in 4.2.3). Then, links between partitions obtained for two adjacent periods can be established by using a distance metric d (e.g. Euclidean or cosine) and an inter-cluster distance measure such as the centroid distance, the single link distance, the complete link distance, the average link distance, or Ward's distance (as defined in Appendix B). Measuring distances between a usage matrix $\mathbf{U}_w^t$ and its adjacent counterpart $\mathbf{U}_w^{t+1}$ yields a distance matrix $\mathbf{D}_w \in R^{K_w^t \times K_w^{t+1}}$, wherein each element corresponds to a pair of temporally contiguous partitions.

This approach has two important drawbacks. First, using inter-cluster distance measures requires an arbitrary criterion for establishing the identity of two temporally adjacent clusters (i.e. below which distance can C_i^t and C_j^{t+1} be considered as the same cluster of word usages?). Furthermore, given a distance matrix $\mathbf{D}_w$, it is unclear in which direction links between partitions ought to be drawn. Indeed, each partition in t can be linked to the closest one in $t+1$ (i.e. $\arg\min_{axis=1} \mathbf{D}_w$) or vice versa ($\arg\min_{axis=0} \mathbf{D}_w$). In the first (forward) case, as every partition in t will have a successor in $t+1$, the disappearance of a usage cluster cannot be detected; specularly, in the second (backward) case, the emergence of new clusters of word use will remain unidentified as every partition in $t+1$ will have a predecessor in t. Selecting the union or the intersection of forward and backward links is equally arbitrary and problematic, yielding contradictory or insufficient inter-cluster connections respectively.

Given the high degree of arbitrariness involved in the above procedure, we turn to a slightly different approach. Instead of independently clustering the time-specific usage matrices $\mathbf{U}_w^1, \dots, \mathbf{U}_w^T$, we perform a single clustering of the contextualised word representations $\mathbf{U}_w$ obtained from all intervals. This results in a total of K_w partitions, which describe the different uses of w in the entire diachronic corpus, as well as in a vector of usage labels $\mathbf{y}_w \in [1, K_w]^N$ (which determines the colouring in Figure 4.1a). The clustering results can then be organised according to the interval of origin of each contextualised representation, thereby producing time-specific vectors of usage labels $\mathbf{y}_w^t \in [1, K_w]^{N^t}$ (which determine the colouring of each bar in Figure 4.1b), where N^t is the number of usages of w in the corpus snapshot of time period t. Normalising by the number of usages we obtain, for each $\mathbf{y}_w^t$, a probability distribution $\mathbf{u}_w^t$ over usage types, illustrated in Figure 4.1c and defined as:

$$\mathbf{u}_w^t[k] = \frac{|\{\mathbf{y}_w^t[i] \in \mathbf{y}_w^t : \mathbf{y}_w^t[i] = k\}|}{N^t} \tag{4.1}$$

This solution comes with multiple advantages. A first, practical benefit is that we circumvent the need for a post hoc definition of identity relations between partitions obtained for different time intervals. A second practical advantage is that the clustering algorithm can better exploit the latent structure of the semantic space spanned by the contextualised representations as (i) it is allowed to build on similarities between usages that belong to different epochs and (ii) it can simply rely on a much larger amount of data points. Yet another, more theoretical advantage has to do with the gradual nature of semantic change: whereas independent clustering runs can result in globally inconsistent partitions, a single clustering of all word uses serves as a form of smoothing and produces more gradual transitions between temporally contiguous periods of word use, which can be crucial when the available data is sparse.

4.4 Quantifying change

The obtained usage type distributions allow qualitative inquiries into the evolution of word meaning. We now propose three main metrics to quantitatively characterise word polysemy and polysemisation: polysemy is expressed as the coexistence of multiple word usage clusters in a single time period whereas polysemisation is described as a change in the relative prominence of usage clusters.

To quantify the degree of polysemy of a word w for which we have obtained the time-specific clusters $C_1, \dots, C_{K_w^t}$, we use the **Boltzmann-Gibbs-Shannon entropy** of the corresponding usage distribu-

tion $\mathbf{u}_w^t$:

$$H(\mathbf{u}_w^t) = -\sum_{k=1}^{K_w^t} \mathbf{u}_w^t[k] \ \ln \mathbf{u}_w^t[k] \equiv -\sum_{k=1}^{K_w^t} p\left(C_k\right) \ln p\left(C_k\right)$$

The entropy $H(\mathbf{u}_w^t)$ of a time-specific usage type distribution measures context-independent uncertainty in the interpretation of a word form in interval *t*. As described in the previous section, however, the clustering step of our procedure produces multiple usage distributions. To quantify how uncertainty over possible interpretations varies across time intervals, we simply compute the difference in entropy between two consecutive usage distributions: $H(\mathbf{u}_w^{t+1}) - H(\mathbf{u}_w^t)$.

An increase in entropy typically but not necessarily corresponds to the generalisation of a word's meaning, while a decrease in entropy typically indicates a specialisation process. Figure 4.2a shows how the word *scene* has been used in an increasingly narrow array of contexts, causing usage A to become vastly preponderant in the 1990s. As expected, to this specialisation corresponds a markedly negative entropy difference (Figure 4.2b). This does not directly correspond to sense acquisition and sense loss, intended as the emergence or disappearance of discrete categories, but rather it indicates a change in the prominence of coexisting usage types, corresponding to the broadening or narrowing of the range of effectively undertaken word interpretations. Broadening and narrowing can be also measured without relying on an aggregation of the word usages, i.e. by simply tracking changes in the variance of the contextualised representations. In practice, this measure of contextual variability has resulted in less precise measurements. When we want to use entropy as a proxy for the degree of semantic change, i.e. regardless of whether it involves generalisation or specialisation, we compute the absolute difference instead: $|H(\mathbf{u}_w^{t+1}) - H(\mathbf{u}_w^t)|$.

As we have mentioned above, however, a difference in entropy does not necessarily correspond to semantic change. Figure 4.2c shows e.g. how usage A is gradually overtaking usage B in the usage distributions of the word *curious*. The difference in entropy between the first two intervals is approximately 0.08, then it decreases slightly for the successive pair of decades: indeed, the first three decades show similar usage distributions. Between 1980 and 1990, i.e. when one usage overtakes the other, the entropy of the usage distributions remains virtually constant, as it can be observed from the drop in entropy difference in Figure 4.2d. This happens as entropy simply measures uncertainty of interpretation, regardless of the ranking of specific usage types.

To take into account not just variations in the size of usage clusters but also *which clusters* have grown or shrunk, we can use measures of similarity between probability distributions. The Kullback-Leibler divergence is often used as a similarity metric but it comes with two disadvantages: it is asymmetric and it requires absolute continuity of the two probability distributions, i.e. the KL divergence is undefined whenever $\mathbf{u}_w^t[j] = 0$ and $\mathbf{u}_w^{t'}[j] \neq 0$. This last property is problematic as it often happens that a usage cluster which is absent in t then appears in t'—such as in the case of the birth of a new usage type. On the other hand, asymmetry is an issue as it is unclear in which direction the KL divergence ought to be measured. To overcome both issues Lin (1991) introduced a variant of the KL divergence, typically referred to as the **Jensen-Shannon divergence** (JSD):

$$\mathrm{JSD}(\mathbf{u}_w^t, \mathbf{u}_w^{t'}) = H\left(\frac{1}{2}\left(\mathbf{u}_w^t + \mathbf{u}_w^{t'}\right)\right) - \frac{1}{2}\left(H\left(\mathbf{u}_w^t\right) - H\left(\mathbf{u}_w^{t'}\right)\right) \tag{4.2}$$

The JSD is symmetric, which allows comparisons between distributions that are not immediately adjacent and makes it irrelevant to establish a direction for the comparison. Even more importantly, the Jensen-Shannon divergence does not require absolute continuity of the two usage distributions—this is crucial to be able to detect the birth and the death of a word sense. Furthermore, JSD can be extended to include n probability distributions (Ré and Azad, 2014):

$$\mathrm{JSD}\left(\mathbf{u}_w^1, \ldots, \mathbf{u}_w^n\right) = H\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{u}_w^i\right) - \frac{1}{n}\sum_{i=1}^{n} H\left(\mathbf{u}_w^i\right)$$

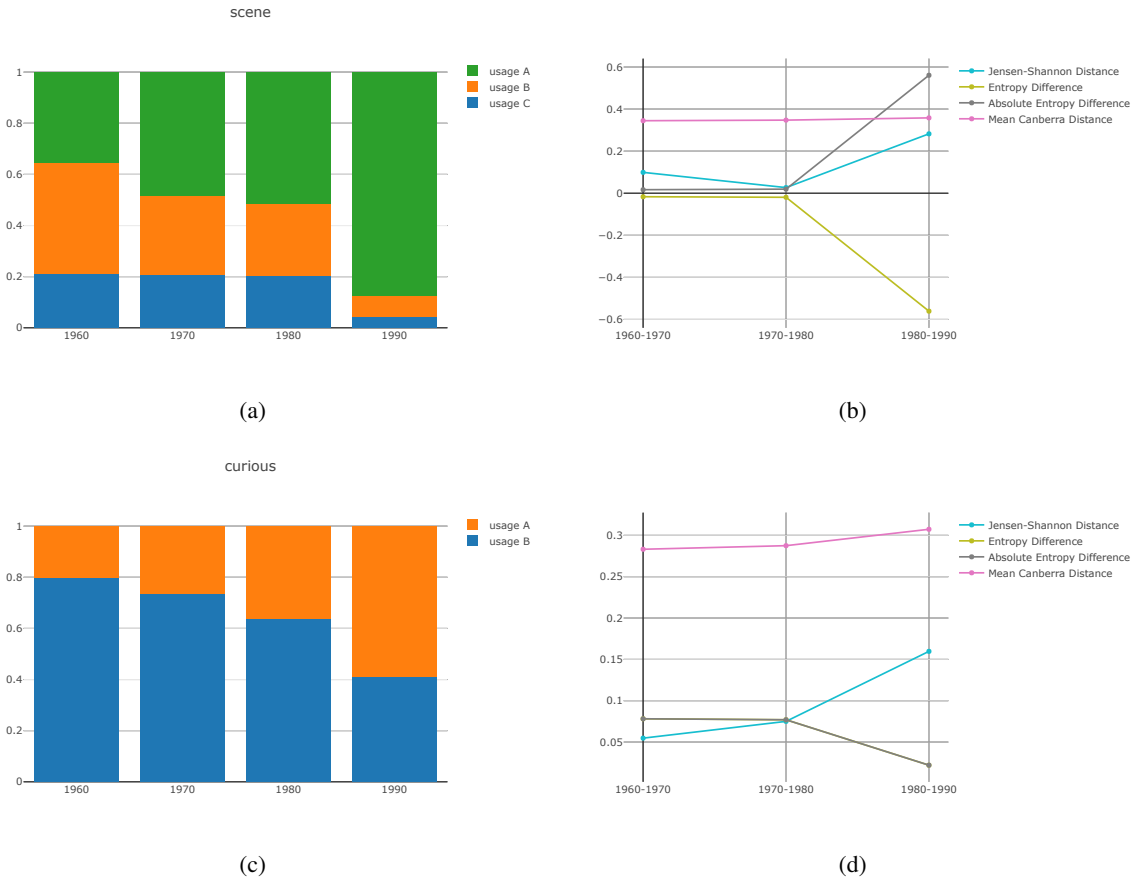


Figure 4.2: Usage cluster distributions obtained with K-Means clustering of contextualised representations of word occurrences from the Corpus of Historical American English (left) and the corresponding quantification of semantic change (right).

In general, very different usage distributions yield high JSD whereas low JSD values indicate that the proportions of usage types are virtually equal across periods; when JSD $= 0$ no shift has occurred. Figure 4.2d shows that, unlike entropy difference, JSD increases between 1980 and 1990 as it is sensitive to changes in the relative predominance of usage types.

Yet another way to quantify change across time periods is to simply measure the **average geometric distance** between usage representations collected in consecutive periods:

$$\frac{1}{N^t \cdot N^{t'}} \sum_{\mathbf{x}_i \in \mathbf{U}_w^t,\ \mathbf{x}_j \in \mathbf{U}_w^{t'}} d(\mathbf{x}_i, \mathbf{x}_j)$$

We experiment with different distance metrics d: Euclidean and cosine distance, as they are mostly used in related work, as well as Canberra distance, a normalised version of Manhattan distance that relies on dimension-wise differences between word vectors and accounts for discrepancies in absolute values across dimensions:

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Intuitively, if every dimension of a contextualised word representation stands for some abstract syntactic or semantic property (or for an unknown aggregation of such properties), one should be able to tell two word uses apart by measuring normalised differences across all such linguistically meaningful (yet hardly interpretable) dimensions. Indeed, Figure 4.2d shows how the average Canberra distance increases

almost imperceptibly between the 1970s and the 1980s and it rises above 0.3 when the predominant usage type has shifted from B to A. Average distance is a stricter metric than the previous two in the sense that a distance value of zero indicates that the word form has been used in exactly the same contexts across time periods. However, as it is an aggregate metric, it can sometimes be less precise: Figure 4.2b shows e.g. how variation in average Canberra distance across intervals may not correspond *in magnitude* to the shift depicted by our usage distributions in Figure 4.2a.

Chapter 5

Evaluation

In this chapter we evaluate the contextualised word representations obtained with BERT as well as our metrics of semantic change. First, we test how the quality of BERT's usage representations varies according to the degree and type of fine-tuning undergone by the language model. Then, we assess the correlation of our measurements of semantic shift with human judgements. For all our experiments, we rely on *Hugging Face*'s implementation of Google's BERT model and choose $\text{BERT}_{\texttt{BASE}}$ as a model variant.[1]

5.1 Fine-tuning

To begin, we present preliminary experiments conducted to understand whether fine-tuning is beneficial to the pre-trained BERT language model. We experiment both with an historical corpus, COHA, and with two conversational data sets, Reddit 2013 and r/LiverpoolFC. The three corpora, as discussed in Chapter 3, present different characteristics. COHA's large size and the heterogeneity of its sources make it representative of general language use. Reddit 2013 is also an heterogeneous dataset but its conversational nature makes it different from BERT's pre-training domain (BooksCorpus and Wikipedia). Lastly, the main characteristics of r/LiverpoolFC is its homogeneity, which makes it representative of the language use of the specific online community that generated it.

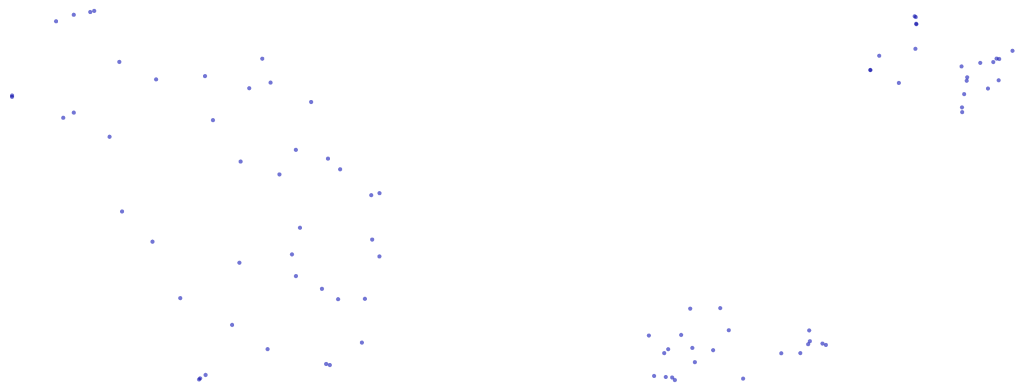


Figure 5.1: T-SNE visualisation of the contextualised representations collected in the r/LiverpoolFC corpus for the word *spicy* with the frozen BERT language model (left) and with diachronically fine-tuned language models (right).

[1] Hugging Face's repository (https://github.com/huggingface/pytorch-pre-trained-BERT) contains PyTorch reimplementations of many state-of-the-art language models, including BERT, which have been tested and shown to match the performances of the original model implementations. Pre-trained models and fine-tuning examples are also available.

Target	Making dumb back *tracking* runs maybe, but Balo doesn't run the channels at all, making him pretty useless for us in our system.
Frozen	Mane is up there with the pressing, energy and back *tracking*.
	Ibe also got ravaged by Can when he picked up that yellow *tracking* back for the errant pass.
	The real *tracking* starts now boys! Before was just a warmup. Now I have all my knowledge on private business jets and can put it to use.
	The flight *tracking* thread will forever be my favourite thread of all time, across every subreddit. Truly hysterical
Fine-tuned	Mane is up there with the pressing, energy and back *tracking*.
	So much back *tracking* haha
	Firstly, Klopp loves hard working, back *tracking* and fast wingers.
	Every movement was stalled once we hit midfield. Just no inventiveness, even Firmino looked poor aside from his back *tracking*.

Table 5.1: A target usage of the word *tracking* and its nearest neighbouring usages, represented by their sentential contexts. Nearest neighbours are determined using cosine distance between representations output by the frozen and the diachronically fine-tuned BERT.

5.1.1 Domain-adaptive fine-tuning

We fine-tune the $\text{BERT}_{\texttt{BASE}}$ language model for $n \in \{1, 2, 3\}$ training epochs and with a learning rate $\eta \in \{0.00003, 0.00001, 0.000003, 0.000001\}$. To test whether domain adaptation to a new target domain is beneficial, we train BERT on the Reddit 2013 conversational data set and measure perplexity on the r/Liverpool dataset. We expect the domain-adapted model to obtain a better perplexity score as it should learn the peculiarities of language use in Reddit conversations. Results show, however, that the frozen BERT outperforms the domain-adapted one across all assessed hyperparameter configurations. This is unexpected (Han and Eisenstein, 2019) but justifiable by the fact that training large Transformer models comes with many difficulties (Devlin et al., 2019; Phang et al., 2018) and it lacks attested procedures (cfr. Section 4.1.3). It is not clear if all layers should be fine-tuned, whether they should be fine-tuned simultaneously or gradually unfrozen, and what the optimal training hyperparameters are.

5.1.2 Diachronic fine-tuning

Next, we fine-tune BERT diachronically using the r/Liverpool corpus, and experiment with a temporal granularity of 3 months, 6 months, and 1 year. The model corresponding to the first interval is initialised with domain-adapted weights obtained training the language model on Reddit 2013. To evaluate the obtained community- and time-specific language models we follow a qualitative approach: first, we choose a selection of words of interest, those distributed and annotated by Del Tredici et al. (2019), then we collect contextualised representations for all usages of the target words, and finally we reduce their dimensionality using Principal Component Analysis (PCA) as well as t-distributed Stochastic Neighbour Embedding (t-SNE). The peculiarity of the usage collection step is that, for each time bin, word representations are obtained with the corresponding time-specific (diachronically fine-tuned) language model.

Although the quality of BERT as a language modeller decreases after domain-adaptation (Section 5.1.1), its quality as a contextualiser seems to increase as a result of diachronic fine-tuning. Indeed, as illustrated in Figures 5.1 and 5.2, the representations of the fine-tuned models exhibit a better separation of different usage types. While this result may appear contradictory (as diachronic fine-tuning follows domain-adaptation), it may be the case that both fine-tuning regimes are too aggressive, thereby producing overfitting effects. On the one hand, overfitting would explain the unsatisfactory perplexity scores of the domain-adapted model, on the other hand, i.e. in the case of diachronic fine-tuning, overfitting can prove beneficial for word sense disambiguation of lexemes with community-specific interpretations (as can be observed in Table 5.1).

All in all, neither fine-tuning procedure seems to produce truly generalisable weights. A possible cause of overfitting may be the large difference in size between BERT's original training corpora (consisting of 3,3 billion words) and the ones deployed in our experiments (all r/LiverpoolFC conversations amount to 40 million words).

5.2 Correlation with human judgements

After this preliminary analysis of the usage representations produced by the frozen BERT as well as its domain-adapted and diachronically fine-tuned counterparts, we turn to an evaluation of our semantic change detection metrics. A quantitative assessment can be performed using lists of words annotated

	Corpus	Pearson's r	Spearman's ρ
Gulordava & Baroni (2011)	Google Books	0.386	n.a.
Frermann & Lapata (2016)	DATE	n.a.	0.377
Skip-gram distance	COHA	0.047	0.119
Entropy difference	COHA	0.217	0.264
Mean distance	COHA	0.224	0.293
Jensen Shannon distance	COHA	0.231	0.224

Table 5.2: Correlation between novelty rankings and human ratings. All correlations are statistically significant ($p < 0.03$) except those obtained for Skip-gram distance.

with a semantic shift index which indicates, for each word, the degree of undergone semantic change according to human annotators. For this analysis, we use the COHA and r/Liverpool datasets in combination with the annotated word lists made available by Gulordava and Baroni (2011) and Del Tredici et al. (2019) described in Section 3.2.

We collect representations for all occurrences of the words of interest using BERT, obtain epoch-wise usage distributions, and then quantify semantic change by relying on the metrics presented in Section 4.4. We experiment with a single frozen BERT model and with diachronically fine-tuned BERT models, as well as with both K-Means and Gaussian mixture models. Only the best results are reported, i.e. those obtained with frozen BERT, K-Means, and silhouette score. Notably, both AIC and BIC proved unreliable for Gaussian mixture models, yielding either the minimum or the maximum number of clusters for virtually every word. As a baseline, we use the cosine distance between type representations obtained for the same word by two incrementally trained Skip-gram models (following the procedure introduced by Kim et al. (2014)). The first Skip-gram model (Mikolov et al., 2013a) is trained on COHA texts from the 1960s and then used to initialise the second Skip-gram model.[2]

Table 5.2 shows the correlation values obtained with our methods together with the scores obtained by previous approaches on the same annotated data set. It should be noted that the correlation values are not directly comparable as the reference models have been trained on different corpora. In particular, Google Books is a much larger collection of texts than COHA or DATE (Frermann and Lapata, 2016); DATE instead is an expanded version of COHA that includes ca. 5 million more words. We use two correlation measures to enable said comparisons, though Spearman's ρ seems to be a better choice given the way the list of words was annotated: whereas Pearson correlation tests for linear relationship between continuous variables, and therefore expects proportionality in the change of novelty and shift scores, Spearman's rank-correlation coefficient tests for monotonic relationships and hence does not expect the two variables to change at a constant rate. This seems to better match the annotation procedure, where judges were asked to assign to each word an index on a 4-point scale (Gulordava and Baroni, 2011). The resulting scores are an aggregation over human annotators hence small decimal differences between scores should not be deemed as particularly meaningful.

The average geometric distance between representations of usages made in contiguous intervals, the difference in entropy and the Jensen-Shannon distance between adjacent usage distributions obtain similar correlation scores, yet lower than those obtained by the competing approaches (cfr. Section 6). We have also experimented with time series that include usage distributions from the intervening decades, the 70s and the 80s, by measuring the correlation of the shift index with the mean and variance of the 4-decade time series (and with the multi-distribution JSD). This has resulted in a decrease in correlation. Finally, we have repeated this qualitative evaluation using a BERT language model that is first domain-adapted to Reddit 2013 and then diachronically fine-tuned on r/LiverpoolFC. The correlation between the output of this model and the semantic shift indices made available by Del Tredici et al. (2019) are not significant, indicating that data sparsity can prove problematic for our procedure.

[2]The Skip-gram models are trained using the `gensim` library: `radimrehurek.com/gensim/models/word2vec.html`. The vector dimensionality is 300, the window size is 5, and the minimum number of occurrences for a word to enter the vocabulary is 5. The model is trained using negative sampling (with 5 samples) as well as downsampling of high-frequency words. All the hyperparameters whose values we have not unspecified are set to `gensim`'s default values.

Figure 5.2: T-SNE visualisation of the contextualised representations collected in the r/LiverpoolFC corpus for the word *spicy* with the frozen BERT language model (above) and with diachronically fine-tuned language models (below). Observations are represented by the sentential context that generated them.

Chapter 6

Analysis

In this chapter we showcase the types of analysis empowered by our method, discuss their usefulness, and scrutinise their mistakes and limitations. Indeed, we believe that the main advantage of our approach is its versatility: a single type of word representation and a single representation learning algorithm can be used to investigate a variety of lexical phenomena. We also discuss potential reasons for a not entirely satisfactory correlation of our shift metrics with annotated semantic shifts, propose a qualitative analysis of usage cluster formation, and elaborate on whether our method is applicable to semantic change happening at any temporal granularity.

The previous chapter presented a quantitative evaluation of our metrics of lexical semantic change. As a (not entirely fair) comparison, we have used distance of collocation-based word representations (Gulordava and Baroni, 2011) as well as the SCAN model (Frermann and Lapata, 2016). The lower correlation scores achieved by our method may result (i) from the quality of the representations, (ii) from an inaccurate aggregation of word usages, or (iii) from the unripeness of our metrics, though we observe that self-distance of Skip-gram vectors (Kim et al., 2014) obtained from our own training data does not yield significant correlation.

With regard to (i), we notice that almost all usage matrices U_w exhibit a high degree of variance, and that trying to reduce the dimensionality of the usage matrices immediately results in a drop of the amount of variance explained by the maintained dimensions. The fact that usage vectors are rather dissimilar to each other suggests that BERT representations are highly sensitive to contextual variations. Although context-sensitivity is clearly a positive characteristic of word features (cfr. Section 2.2.2), excessively high sensitivity may make it more difficult to obtain meaningful aggregations of usages. In other words, the interpretation obtained e.g. for the word *free* in sentence (1-a) should not differ much from the interpretation obtained from sentence (1-b), where a single distant word has changed, or from sentence (1-c), where multiple tokens have changed—even one that is adjacent to *free*. The meaning in context of *free* is the same for all example sentences.

(1) a. He was not tased because of his viewpoint or to restrict his *free* speech; it was because he would not yield the floor to other students.
b. He was not tased because of his viewpoint or to restrict his *free* speech; it was because he would not yield the floor to other **employees**.
c. **She** was not tased because of **her** viewpoint or to restrict **her** *free* speech; it was because **she** would not yield the floor to other students.

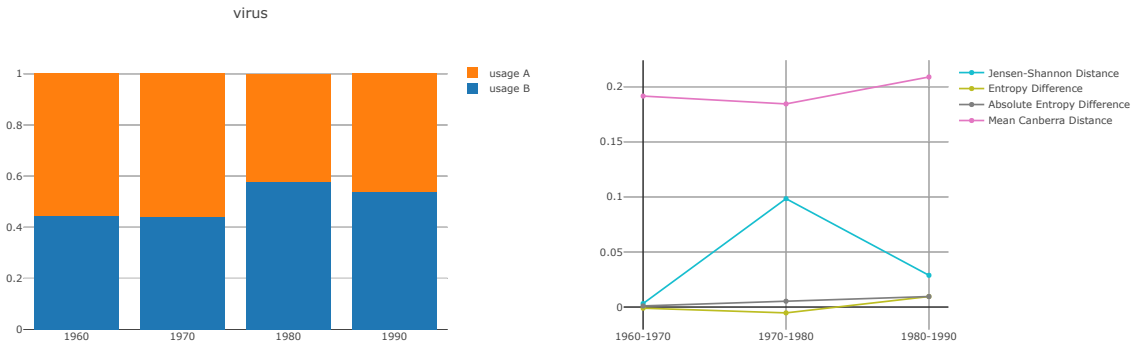
To define the meaning of a target word, distributional word representation models take into account, with uniform importance, a fixed number of surrounding tokens. The above examples[1] show, however, that drawing a boundary for context-sensitivity is not a straightforward endeavour, and that it is preferable to let the representation learning algorithm set such limit automatically, dynamically, and gradually, as BERT does. So high sensitivity is necessary and only constitutes an issue if the model's attention is

[1] Sentence (1-a) is taken from COHA, document `news_2007_641837.txt` whereas sentences (1-b) and (1-c) are constructed by the author.

drawn by irrelevant cues. Fortunately, the latter does not seem to be the case as very recent studies show that while *single* BERT attention heads do not address multiple linguistic relations, specific heads learn to find direct objects of verbs, determiners of nouns, objects of prepositions, and objects of possessive pronouns with at least 75% accuracy (Clark et al., 2019).

Given the rather high quality of BERT's contextualised word representations, the weak results of our quantitative assessment may be caused by an inaccurate aggregation of word usages (ii). The next section of this chapter discusses this possibility at length and concludes that most usage aggregations obtained with our method are interpretable and meaningful yet that some of them may not correspond to those a human reader would produce.

Finally, concerning (iii), we interpret the fact that the three proposed metrics achieve similar correlation scores as an indication that they offer complementary accounts of transformations in usage distributions. A change in entropy between temporally contiguous usage distributions indicates a process of narrowing or widening of the possible interpretations that a word form may undertake. However, a lexical semantic change is often accompanied by yet does imply such a change in uncertainty of interpretation (cfr. Section 4.4). On the other hand, although the Jensen-Shannon distance between two distributions should be a more direct measure of actual change in the referential scope of a term, it also yields positive values when small oscillations occur in the relative predominance of a word usage type (Figures 6.1a and 6.1b), which should not be necessarily deemed as a sign of shift. Lastly, although it does not rely on usage distributions, the average geometric distance between temporally adjacent usage matrices is also the result of an aggregation, namely of computing the mean of all pairwise distances. A possible solution to this problem may have come from computing distances only between time-specific cluster centroids, but additional experiments have shown that this is not the case.



(a) (b)

Figure 6.1: Usage cluster distributions obtained with K-Means clustering of contextualised representations of word occurrences from the Corpus of Historical American English (left) and the corresponding quantification of semantic change (right) for the word *virus*.

6.1 Cluster formation

The state-of-the-art results obtained by BERT on many token-level language tasks are an indicator of the good quality of BERT's contextualised word representations. Indeed, as explained in Sections 2.2 and 2.3, we extract word features from the BERT language model as we expect them to encode those collostructural word properties that are intrinsically ignored by standard distributional approaches. To establish if BERT's dynamic interpretations of word usages are indeed more discriminative and thus more informative than static type-based representations, we analyse the linguistic properties encoded by the model. We do so by investigating which properties are shared by usages that are clustered into the same partition.

As a point of departure, we observe that the contextualised representations obtained for polysemous words tend to cluster according to the respective underlying senses of those words. As an example, the

vectors collected for the word *curious* are grouped together depending on whether *curious* is used to describe something that excites attention as odd, novel, or unexpected, or rather to describe someone who is marked by a desire to investigate and learn. As shown in Figure 6.2a, both usages are present across the two centuries under scrutiny: the cluster of usages A is formed by occurrences such as: *full of questions, intensely* ***curious*** *and entirely non threatening* or *staring at him, half fearful, half* ***curious***, whereas cluster B consists of usages such as *the most* ***curious*** *reading* or ***curious*** *sense of gratitude*. Similarly, occurrences of the word *users* are discriminated according to whether they refer to users of digital products and services (usages B and E: *computer* ***users***, ***users*** *of both Yahoo and AOL*), users of resources (usage D: ***users*** *of hydro-electric energy*), users of non-digital products (usage F: *car* ***users***), drug users (usage C: *dealers and* ***users***; *many* ***users*** *quit on their own*; *pill* ***users***), or very specifically to the Suez Canal Users' Association, which is granted its own cluster (usage A: *the* ***users'*** *association*).

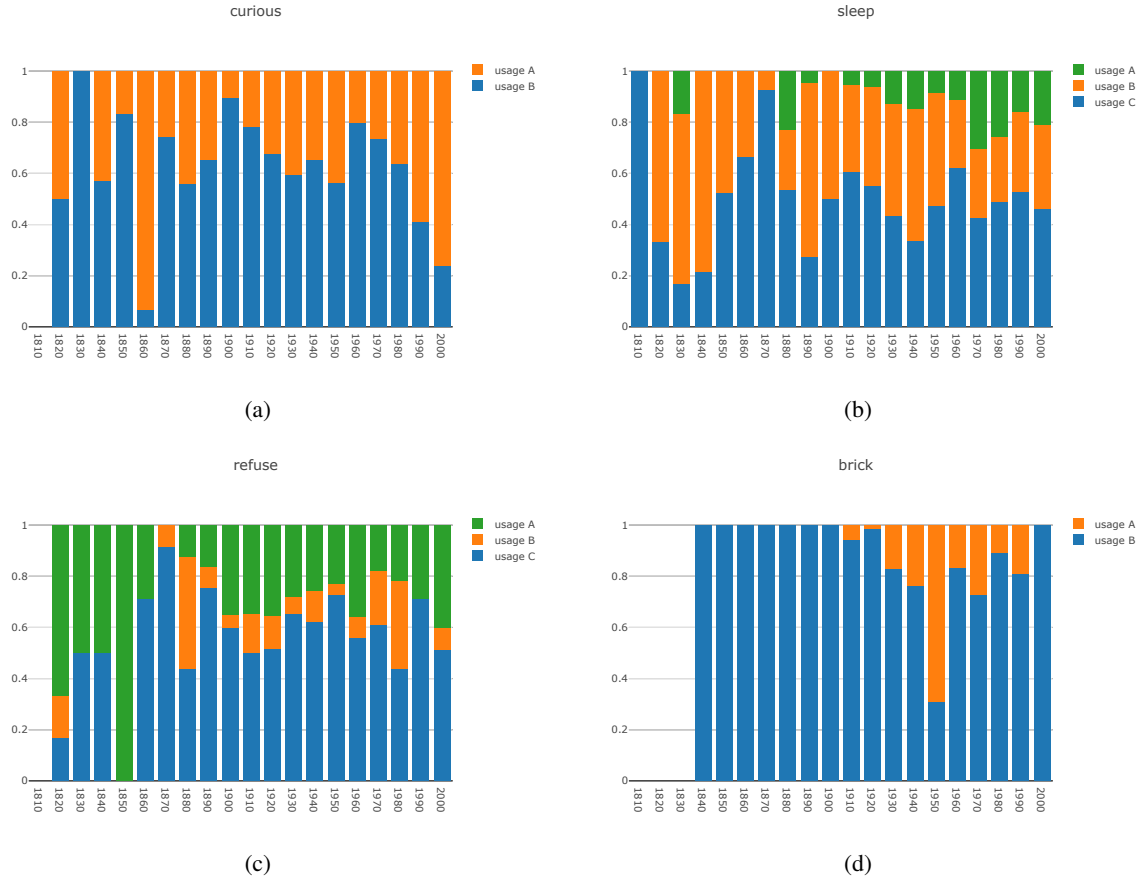


Figure 6.2: Usage cluster distributions obtained with K-Means clustering of contextualised representations of words from the Corpus of Historical American English. Specific usage types of each word are described in Section 6.1.

Contextualised representations also seem to discriminate literal from metaphorical usages. Usage vectors for *ceiling* form two clusters: one corresponds to references to ceilings as the upper interior surface of rooms such as *the* ***ceiling*** *of a church* or *those who prefer the open sky to a* ***ceiling***, while the other cluster corresponds the word's metaphorical sense of upper limit, usually related to money (***ceiling*** *prices*; *a tax-****ceiling*** *amendment*), even when the surrounding sentence involves cues to the literal sense, as in *to keep from breaking through the* ***ceiling*** *the treasury has already suspended the sale of its savings notes*. The same is true for *sphere*, which is interpreted either as a round solid figure (*to make the perfect* ***sphere***; *at every point of this* ***sphere*** *it tends toward the centre of it*; *the whooshing* ***sphere***) or as an area of knowledge, activity, interest, etc. (*the* ***sphere*** *of immaterial goods*; *many subjects fall within its* ***sphere***; *it is the United States to enlarge the* ***sphere*** *of its action*).

As expected, contextualised word representations do not only encode semantic relatedness or similarity, they also enable differentiation based on a word's syntactic functionality. Part-of-speech ambiguity, for example, appears to be at least partially solved by the language model. Occurrences of *cost* are partitioned according to whether the word is used as a noun (A: *the* ***cost*** *of the materials*) or as a verb (B: *how much they will* ***cost***), and regardless of the part of speech of the surrounding words: cluster A also includes *a further* ***cost****-price squeeze* and *a major* ***cost*** *component of aluminum* while cluster B contains usages such as *will [...]* ***cost*** *5 less per bottle* or *they* ***cost*** *even more*. The same holds e.g. for *excuse*: one cluster contains the word's verbal use such as in *I will not* ***excuse*** *you* or *having to* ***excuse*** *herself*, whereas the other partition corresponds to the nominal use, as in *this* ***excuse*** *is not good* or *his want of knowledge was no* ***excuse***.

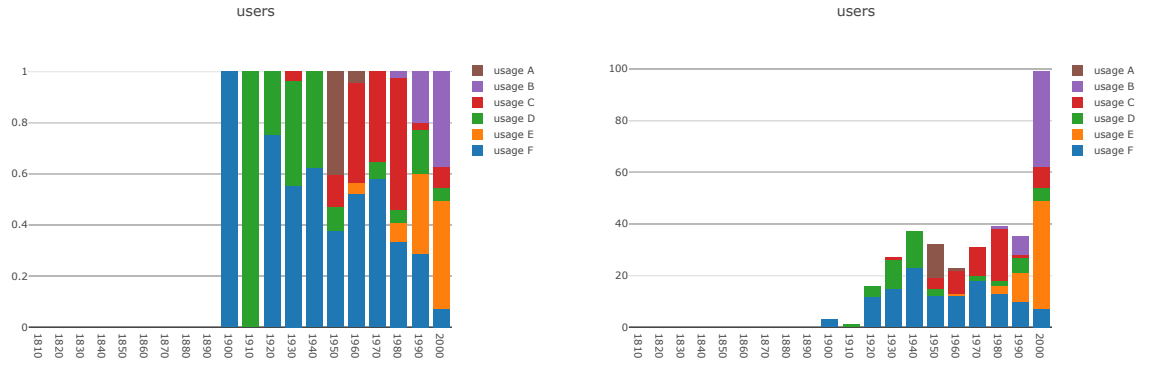
Furthermore, and perhaps less expectedly, BERT pays attention to the presence and type of syntactic arguments that usually co-occur with a lexical item. E.g. usages of the word *refuse* are clustered together according to what seem to be its subcategorisation frames (Figure 6.2c): partition A contains occurrences of *refuse* used either as an intransitive verb or as a transitive verb followed by a direct object (***refuse****, and you die*; *he would not* ***refuse*** *a draft*), cluster B contains mostly nouns (*the* ***refuse*** *of the schools*), and cluster C contains verbs with infinitive complementation (***refuse*** *to hire*). Interestingly—and a sign that BERT is able to cope with distant dependencies—intervening tokens do not seem to distract the language model: the usage in *can railroad corporations* ***refuse*** *or neglect to perform their public duties upon a controversy* belongs to cluster C even though *refuse* and its syntactic argument *to perform* are not linearly adjacent. Other examples of same-POS occurrences that are discriminated by our model include verbs and nouns used as modifiers. The word *family* is recognised either as a noun (***family*** *and friends*) or as a nominal modifier (***family*** *members*), and *dining* is interpreted either as a verb (***dining*** *at tables*; *immediately after* ***dining***), or as a verbal modifier (***dining*** *facilities*, ***dining*** *and living rooms*). In fact, *dining*'s verb partition also includes substantivised verbs (*extravagant wining and* ***dining***), suggesting that the features onto which our model latches may not always be predictable (i.e. why are substantivised forms not assigned to a separate partition?). Singling out modifiers of noun phrases, however, seems to be a recurring strategy of our model. We see this behaviour also for the word *sleep* (Figure 6.2b), whose usages are partitioned into nouns (usage B: *a good night's* ***sleep***), verbs (usage C: *you'll feel and* ***sleep*** *better*), and modifiers (usage A: ***sleep*** *habits*; ***sleep****-wake cycle*).

Specific partitions are also assigned to lexical items that are used to refer to entities, so that *mirror* can be a polished surface that forms images by reflection (*a sink and a* ***mirror***) as well as the name of a news outlet (*bought by The Times* ***Mirror*** *Co.*; *the tabloid* ***Mirror***; *the Los Angeles* ***Mirror***[2]). This is the only differentiation the model applies to *mirror*: metaphorical nominal usages as well as verbal usages are included in the first cluster. The latter example hints again at the partial unpredictability of our model when it comes to choosing the lexical and contextual properties that define usage clusters. A more surprising instance is the word *doubt*, for which we do not observe the expected noun-verb distinction but rather a distinction between usages in affirmative contexts (*there is still* ***doubt***; *the benefit of the* ***doubt***), in negative contexts (*there is not a bit of* ***doubt***; *beyond a reasonable* ***doubt***), and usages that are modified by one or more morphemes (***doubt****less*; ***doubt****ingly*). Why are such usages even counted as occurrences of the word *doubt*? The reason is that BERT's tokeniser splits actually occurring tokens into Wordpiece tokens (Section 4.1.1), so that the model reads *doubtless* as *doubt ##less*. This peculiarity of BERT sometimes causes the emergence of unexpected clusters of proper names or of mixed referents, mostly in the case of virtually unambiguous words such as *lips* (e.g. *John Lipsky* / *john* ***lips*** *##ky*), *brick* (e.g. *Marshall Brickman* / *marshall* ***brick*** *##man*), and *card* (*cardiorespiratory* / ***card*** *##ior ##es ##pi ##rator ##y*, *cardamom* / ***card*** *##amo ##m*). As can be observed in Figure 6.2d, this is not always a neglectable issue.

Another consideration is to be made with regard to the clustering algorithm that produces the types of usage differentiation described in this section. As K-Means and its K selection criteria allow for a minimum of two clusters (cfr. Section 4.2.3), our model is sometimes forced to build multiple partitions for unambiguous words, according to hardly interpretable features. This happens e.g. for the words

[2]Capitalisation is used here only to improve readability. All words are seen by our model in lower case.

maybe, *woman*, and *women*. Finally, we notice that the cluster distributions obtained in the first decades (typically until 1870) are in general less reliable than those obtained for later intervals. This seems to be due to the limited amount of word occurrences found in those periods, which is directly related to the overall smaller size of the text collections available for earlier periods (cfr. Section 3.1.1). Indeed, words that occur rarely across all decades (roughly, less than 300 times *in total*) exhibit less interpretable usage distributions.



(a) (b)

Figure 6.3: Usage type distributions and frequency distributions obtained with K-Means clustering of contextualised representations of the word *users*, as it occurs in the Corpus of Historical American English. Specific usage types of each word are described in Sections 6.1 and 6.2.

6.2 Lexical change modelling

The usage distributions we obtain by clustering contextualised word representations give us a way to pinpoint the exact interval where a certain word usage has first appeared, its last interval of occurrence, as well as the overall diachronic stability of a cluster. In particular, we declare the *death* of a usage cluster in interval t', when the probability of the corresponding usage type is 0 in all intervals $t \geq t'$. Similarly, the *birth* of a cluster happens at time t' when the corresponding cluster is empty for all intervals $t < t'$. The *stability* of a cluster is measured as the proportion of intervals wherein the corresponding usage type has occurred. These measures will help us describe the trajectories of the set of words under scrutiny.

As an example, if we look at the usage distributions of the word form *users* (Figure 6.3a), we notice, first, that the word never occurs before 1900. Then, our visualisation shows that the initial occurrences of *users* refer exclusively to users of resources and products (usages D and F); these two usage types are the most stable across decades, with stability values of ca. 82% and 91% respectively. We can further observe the birth of four usage types: people making use of narcotics have been called *users* starting from the 1930s (usage C: *Pot prohibition gives sporadic* ***users*** *the stigma of criminal records and makes young people cynical*), users of digital products and services (usages B and E) have been designated by this word since the 1980s. The fourth birth, in the 1950s, corresponds to occurrences of *users* that refer to the Suez Canal Users' Association (usage A); this usage dies in the 1960s and its stability can be quantified as ca. 18%. The careful reader will have noticed that the birth of cluster E is actually located in the 1960s. We have stated otherwise as there exist only a single usage of type E in the 1960s, which we consider to be assigned to the incorrect partition (cfr. Table 4.1). How can we facilitate such diagnoses?

Recall that our epoch-specific probability distributions result from normalising frequency distributions (Equation 4.1). By dropping the normalisation, we can combine information about the relative and absolute frequency of co-occurring usage types as well as the overall frequency of the word form. Figure 6.3b shows e.g. that cluster E is of negligible size in the 1960s, that users of resources are mentioned less frequently starting from the 1950s and slightly more often in the last two decades (perhaps as a result of increased attention towards excessive use of natural resources), that deployment of *users* in the context of digital products and services is scarce in the 1980s, it increases to ca. 50% of all usages in the 1990s,

and it explodes in the 2000s, when it is not only responsible for 80% of occurrences but also for a very large growth in the overall frequency of the word form—with this new set of referents, the word *users* has become much more widespread.

Such results show that, as expected, our method is able to detect polysemisation processes. In particular, it is able to recognise the broadening trajectory of a word's interpretation that is driven by specific events, technological innovations, and cultural transitions. An example of event-driven broadening can

Figure 6.4: Usage type distributions obtained with K-Means clustering of contextualised representations of words occurring in the Corpus of Historical American English. Specific usage types of each word are described in Section 6.2.

be again observed in Figure 6.3b, where usage A corresponds to a very specific usage of the word *users* within the named entity Suez Canal Users' Association. The corresponding cluster survives for two decades and is mostly prominent in the 1950s, exactly when the Suez Crisis (or Second Arab-Israeli

War) took place. The word *curtain* too acquires a metaphorical meaning within the locution *iron* ***curtain***. Unsurprisingly[3], the corresponding usage cluster B pressingly emerges in the 1940s, it takes up to 62% of the word usages in the 1950s and it appears for the last time at the end of the cold war, in the 1990s (Figure 6.4c). Another war-related example is the word *atom*, whose usage in phrases such as ***atom*** *attack*, ***atom*** *bombs* (cluster B) appears in the 1930s, grows both in relative and absolute frequency in the decades around World War II, and then decreases again, with a trailing drop in overall frequency (Figure 6.4a).

Many detected semantic shifts, on the other hand, are driven by technological innovations. Examples of these shifts are the words *virtual*, *lift*, and *energy*. The word *virtual* occurs intermittently from the 1840s, and for a century it only designates the property of being something in effect though not formally recognised; in Figure 6.4b this corresponds to cluster B, including usages such as ***virtual*** *dictator* or ***virtual*** *monopoly*. Starting from the 1950s, the word *virtual* begins to be used in different types of contexts (e.g. ***virtual*** *particles*) until, in the 1990s and 2000s, it acquires the meaning of something that does not physically exist but appears to do so thanks to e.g. a computer software; this is usage A: ***virtual*** *private networks*; ***virtual*** *guitar, keyboards or drums*; *a* ***virtual*** *walk*. Similarly Figure 6.4e shows how, in concomitance with the availability of automated lifts, at the beginning of the twentieth century the word *lift* starts referring to platforms or compartments for raising and lowering people or things to different levels. Usages of the word *energy* are clustered by our method into two partitions, as illustrated by Figure 6.4d. One corresponds to the sense of strength and vitality as well as its metaphorical extension to the domain of physics as a property of matter and radiation (usage A: *a drop in* ***energy*** *in the afternoon*; *the amount of matter and* ***energy*** *in the universe*) and it remains predominant until the 1930s; the other contains interpretations of *energy* as usable power derived from physical or chemical resources such as heat or electricity (*industrial* ***energy*** *consumption*; *the atomic* ***energy*** *commission*) and it never represents less than 25% of the word occurrences starting from 1940.

Cultural transitions too can cause words to appear in new contexts. Our model detects e.g. the metaphorical extension of the word *mobility* to the domain of society (*social* ***mobility***; *upward* and *downward* ***mobility***; *middle-class* ***mobility***) or the gradual shift of *coach*, shown in Figure 6.4f, from referring to vehicles (usage A: *a two-door* ***coach***; ***coach*** *round trips*; *the fairy godmother changes a pumpkin into a* ***coach***) to designating trainers (usage B: *a teacher or* ***coach***, *basketball* ***coach***).

In sum, our model is able to detect semantic broadening, narrowing, as well as metaphorisation. We in fact maintain that it can also identify metonymy as metonymical usages typically show very different collostructural properties compared to their literal counterparts. Furthermore, pleasantly in line with linguistic theories of change (Section 2.1), our approach models semantic change as a gradual process of polysemisation: a shift from a word sense A to a new sense B never occurs directly but rather through intermediate polysemous stages (see e.g. Figure 6.4f).

6.3 Temporal granularity

We also experiment with representations obtained using a corpus of finer temporal granularity in order to establish whether our method is sensitive to semantic change that happens at a faster pace. We use COCA, which ensures that our method has a sufficient amount of observations for each year—virtually the same amount of texts available for each decade of the COHA dataset (cfr. Section 3.1.1). Excluding this possible confounding factor, what we evaluate is the model's ability to distinguish among a generally less diverse variety of usages (as they are produced in a shorter time span) and thus to detect more subtle changes in the usage of a word. As, to the best of our knowledge, there exists no annotated list of words changing meaning in the last three decades, we collect words which have been used by Davies (2010) as examples for the semantic shift analysis empowered by the Corpus of Contemporary American, and we include a few hand-picked words from the evaluation set provided by Gulordava and Baroni (2011).

First, we examine the cluster formation rules followed by our method and we note that it is able

[3]On 5 March 1946, at Westminster College in Fulton, Missouri, Winston Churchill's used the term "iron curtain" in his so-called Sinews of Peace address: *From Stettin in the Baltic to Trieste in the Adriatic an iron curtain has descended across the Continent. Behind that line lie all the capitals of the ancient states of Central and Eastern Europe.*

to detect meaningful variation in usages also at this finer granularity. As it did with texts in COHA, our method can discriminate between underlying senses of polysemous words (e.g. *monitor*, *virtual*), between usages that fulfil a different syntactic functionality (e.g. *download*), as well as between literal and metaphorical word usages. It also recognises usages that are part of named entities (e.g. *Verizon* ***Wireless*** *Theater*, usage C in Figure 6.5b) or of phrasal collocations (e.g. *global* ***warming***), and it constructs clusters for usages of words as morphemes within larger lexical items (e.g. ***wireless****ly*). As we observed in Section 6.1, our method does not follow generalisable rules: e.g. the nominal and verbal uses of *book* are not separated. Indeed, as the number and nature of the usage types are completely data-driven, we may not observe clusters that relate to what sometimes seem to be obvious underlying senses or, on the contrary, we may observe unexpected clusters.

Next, we look into the different types of change modelled by our approach and find that they largely overlap with those detected in COHA texts. This is a positive result as it indicates that the quality of our method does not depend on the temporal granularity of the available data sets—as long as the amount of observations is sufficiently high (cfr. Section 6.2). Among the detected cultural drifts, we can again detect changes driven by technological innovations, important events and entities, and we observe how frequency distributions mirror correspondingly expected increases and decreases in word use. We also find linguistic shifts which, at this finer granularity, seem to address new communicative needs brought about by technological and cultural innovation.

As a first example of cultural drift, Figure 6.5a shows that our method detects two usages of *warming*. One cluster has remained steady across the years; it is the one that includes adjectival and verbal occurrences of the word, describing the property or the act of causing an increase in the temperature of something (usage B: *my guy and I decided to try something new in bed, so I bought* ***warming*** *oil and gave him a full-body massage*; *have tomato sauce* ***warming*** *in a large pan*). The other cluster corresponds to usages within the phrase *global warming* or related to global warming as a topic (usage A: *nitrogen is now understood to help regulate the carbon cycle and exert both cooling and* ***warming*** *effects on the climate*). The frequency of such usages has largely increased starting from 2000 and it has become predominant. This is an example of a pure cultural drift, one that is not determined by a specific innovation or event.

Some cultural drifts are instead related to particular events (e.g. *crisis*), entities (e.g. *wireless*), or technological advances—more (e.g. *web*, *download*) or less directly (*reality*, *virtual*). A dangerous moment or a time of great confusion can be referred to as a *crisis* (usage B: *taking command of the* ***crisis***; *sudden* ***crisis*** *or emergency the* ***crisis*** *of values*). The term is also used when difficult situations involve society at large as well as globally relevant systems and institutions (usage A: *from* ***crisis*** *to stagnation*; *it was government interference in them that caused the* ***crisis***). As shown in Figure 6.5c, while the frequency of usage type B exhibits only small fluctuations across the years, the frequency of usage type A increases in concomitance with famous systemic crises: *the* ***crisis*** *of communism* in 1990-1991 and the financial crisis begun in 2008 (*a debt* ***crisis***; *the government's anti-****crisis*** *plan for 2009*). Another intriguing example is *reality*, whose evolution is depicted in Figure 6.5f. The term, in its most prototypical sense, refers to the state of things as they truly are (usage B: *young children have trouble distinguishing between* ***reality*** *and fantasy*) yet an increasingly large proportion of usages (type A) refer to *tv* ***reality***, occur within the phrase *virtual* ***reality***, or they are related to it (e.g. *online* ***reality*** *game*).

The linguistic shifts detected by our procedure are a case in point of how language varies according to the communicative needs of speakers. In Figure 6.5b, we can observe e.g. how the percentage of uses of *wireless* suffixed by adverbial morphemes (usage A: ***wireless****ly*) increases with respect to adjectival uses (usages B and D; usage C corresponds to the named entity Verizon Wireless Theater). Until the 2000s, *wireless* was almost exclusively used as an adjective, hence without affixation, to describe the property of specific referents not to need wires to establish a connection (***wireless*** *device*; ***wireless*** *networks*). From the 2000s (and with the exception of a dubious peak in 1994), adverbialised usages have begun to increase. We can interpret this use of morphologisation for adverb formation as the result of the increasingly frequent need to express that *any action* has the quality of not needing wiring in order to be performed: what used to be the property of particular referents has become a general category in

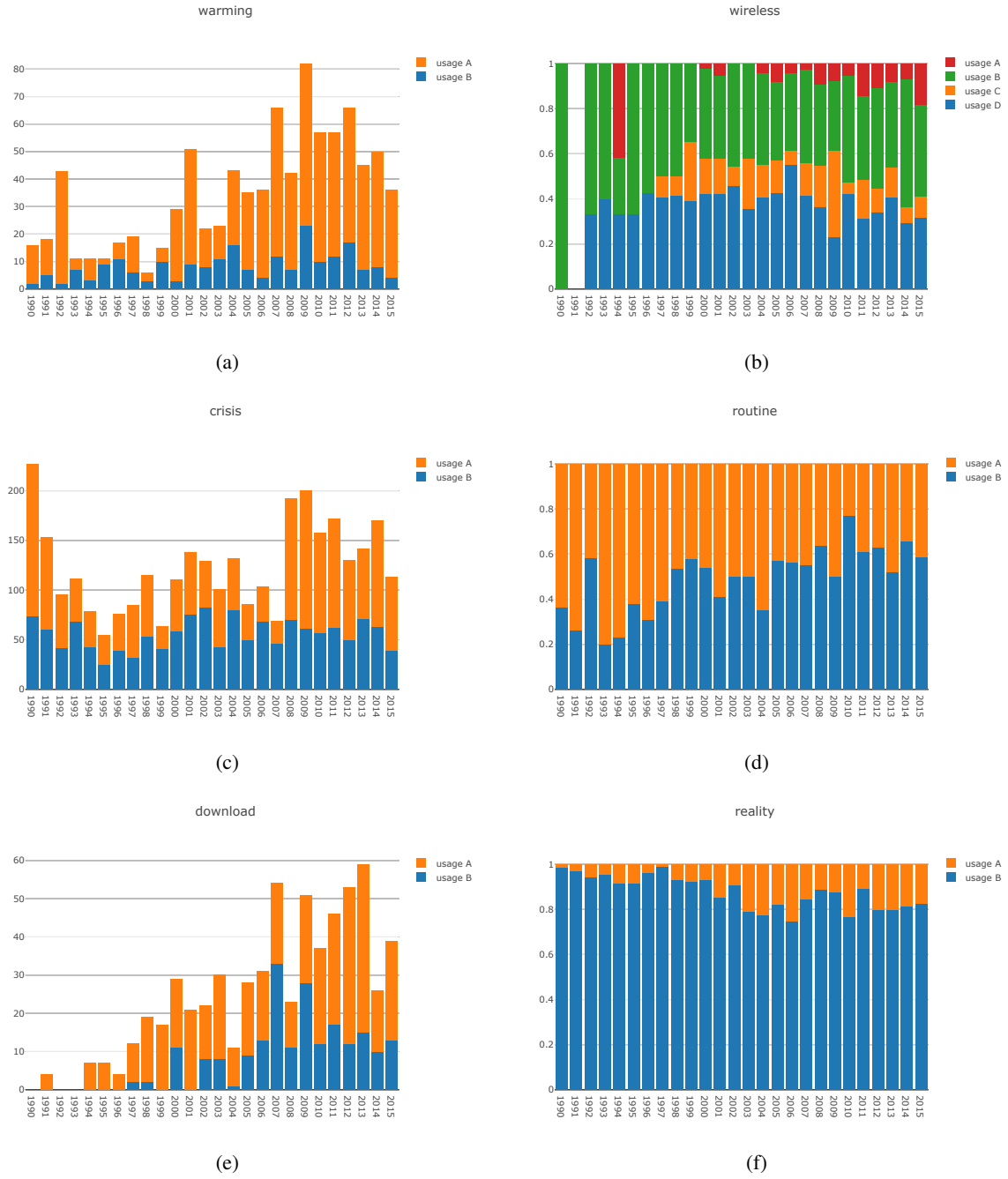


Figure 6.5: Usage cluster distributions obtained with K-Means clustering of contextualised representations of words from the Corpus of Contemporary American English. Specific usage types of each word are described in Section 6.3.

the English vocabulary. The words *routine* and *download* are similar examples. Figure 6.5d shows that *routine* has increasingly been used nominally (usage B), to designate a usual way of doing things consisting typically of a fixed set of activities, rather than as an adjective (usage A) to describe specific referents as ordinary and not special. Likewise, Figure 6.5e depicts how *download* used to be employed as a verb to express the newly available action of transferring data from a server to a device (usage A: e.g. ***download*** *updates*). In the twenty-first century, is being used as a noun too, becoming something countable and qualifiable (usage B: e.g. *free* ***download***).

These small case studies exemplify how new semantic affordances carried by innovation become lexified as soon as the language community shares a recurring need to verbalise them. As we hoped, trying

to understand the processes that lead to linguistic change provides us with more general insights into how language is used to create socially recognised meaning.

Chapter 7

Conclusions

The meaning of words is in constant change. If, in the past, linguists had to read through thousands of books in order to detect lexical semantic change, nowadays we can rely on automated distant reading of digitised books as well as on increasingly large collections of published texts, scripts, and everyday language. The distant reading programme is motivated by a will to go beyond the thousands of words that a single individual is able to read, and take off to observe the language production of entire communities from above—a bird's-eye view free from the subjective biases involved in close-reading interpretation. To do so, language scientists have been making an effort to develop computational techniques for the automatic analysis of raw language data: they have begun by counting word collocates, used the latter to derive abstract high dimensional word representations, and then embedded words in high dimensional semantic spaces with the use of predictive neural networks. We now propose to use language models as a tool for representing words as functions of their collostructural properties and to use the resulting word features for the tracking and analysis of diachronic lexical change.

Together with the evolution of computational models of semantic shift, the theoretical underpinnings of such automated methods have evolved too. The first contribution of this thesis is therefore a review of different word representation models and of their application to lexical semantic change modelling. Starting from type-based representations, we have discussed the disadvantages of modelling the meaning of a word using static features which are applicable to all occurrences of that word. Our review has then moved from sense-agnostic word features to the sense-aware modelling of word types as the mixture of multiple concurring underlying senses (Tahmasebi et al., 2018). These approaches account for polysemy by discretising a large variety of word usages into a fixed number of senses, and they come with three main downsides. First, they require arbitrary assumptions concerning the expected degree of polysemy of a word (how many senses does w possess? How many of those will be observed in a given dataset?). Second, they produce context-independent word features and third, they do so by largely ignoring the meaning carried by grammatical and rhetorical structures. To address these limitations we have proposed using neural language models to produce usage representations, i.e. contextualised word representations that define every single word usage as a function of its context of occurrence.

The second, larger contribution of this thesis is a method for the aggregation of said contextualised representations into meaningful clusters of word usages. We demonstrate the interpretability of the approximate rules which guide cluster formation and rely on the resulting partitions to obtain time series of usage type distributions. The cluster formation procedure is entirely data-driven, with the goal of excluding subjective biases from the modelling process, and the resulting usage type distributions are characterised by actual word usages from the corpus thus particularly apt for qualitative analyses of semantic shift. Nevertheless, to quantify over these distributions and over collections of contextualised word representations, we propose three metrics of semantic change and evaluate them against two human-annotated data sets. Although our measures exhibit a weaker correlation with human judgements than previous attempts (Gulordava and Baroni, 2011; Frermann and Lapata, 2016), qualitative analyses demonstrate that our method is able to detect narrowing and broadening trajectories as well as metaphorisation and, potentially, metonymisation. The identified polysemisation processes are cultural drifts—lead by specific events, technological innovations, and cultural change—as well as linguistic shifts, such as modifications

of the subcategorisation frames of nouns and verbs.

A further goal of this thesis is to show that language models and contextualised word representations are versatile tools for the analysis of language change and variation in general. Indeed our approach offers a double perspective on word meaning which is the result of combining the semasiological and onomasiological views: we go from word form to function in that our *unit of analysis* are lexical items, and from function to form as our *unit of representation* are the functions fulfilled by those lexical items—observed directly from the data.

Our work is not a comprehensive collection of the types of inquiries that contextualised word representations allow. Yet we hope that, by demonstrating their wide applicability and ease of use, this thesis will spark interest in developing further usage-based methods, refining current language representation learning models, as well as using them in practice to answer linguistic and sociolinguistic questions.

Naturally, our method comes with drawbacks. For instance, if standard methods find it problematic to reliably and accurately model low-frequency words, our approach still does not fully solve this issue. Another limitation of our method is that it analyses the change of a possibly large but still limited amount of lexemes. In other words, our method does not allow to simply detect changing words over the entire vocabulary of a corpus. This would either require storing a high-dimensional contextualised word representation for every single token appearing in the corpus, or following the more approximative strategy described in Appendix C.

While these are the limitations of contextualised word representations for lexical semantic change modelling, we believe that they are remarkably outweighed by the potential of such representations, which provides inspiration for future research. Firstly, the metrics that we have defined result in time series which have not been fully exploited in the current work. Time series analysis can reveal the exact change point of a semantic shift (as shown e.g. by Kulkarni et al., 2015) with the precision granted by statistic significance measurements. Moreover, it allows for the analysis of trend and seasonality of word usages (e.g. the word *tracking* is used in r/LiverpoolFC in its sense of tracking a football player in order to have him sign a contract; this usage type typically re-occurs every year during transfer seasons). Secondly, considering that token- and sentence-level representations output by pretrained language models have proved to be good features for tasks such as sentiment analysis, we expect usage representations to be also employable to characterise more types of semantic change, such as amelioration and pejoration.

As previously mentioned, contextualised representations are not limited to an analysis of diachronic change but they can also be applied to synchronic studies of semantic variation: the time-stamped usage distributions that we have presented can be easily extended to include a third dimension of variation—e.g. genre, social status, geographical position, or political orientation. As an example, it is reasonable to expect that usages of the word *significant* vary from academic texts, to books of fiction, to newspapers. Moreover, contextualised representations allow for a fully onomasiological approach which can prove particularly useful e.g. for synonymy detection: moving from function to form, words with overlapping semantic affordances (Szymanski, 2017) may be identified by finding regions of the semantic space that are occupied by usage vectors of different lexemes. This paradigm can be further extended to a multilingual setup (Beinborn and Choenni, 2019), where multiple languages are represented in the same semantic space. Following the lexical typology tradition, semantic maps can then e.g. be constructed using contextualised representations as nodes and finding relations between word forms of different languages based on the overlap of their usage representations.

The application of semantic variation and change modelling are only limited by the creativity of researchers and professionals. In their review, Kutuzov et al. (2018) mention two broad categories of possible applications: on the one hand, linguistic inquiries into the dynamics and underlying causes of semantic shifts, and on the other, language-based event detection approaches. Social scientists may try, in the furrow of this work, to define the underlying *mechanisms* of language change, i.e. to define a set of rules that can explain and predict the emergence and disappearance, in a language community, of new meanings given a word form, as well as of new symbolic ways of expressing already existing semantic affordances. In addition, criminologists may compare the linguistic distribution of coded terms with that of candidate translations to decipher the intended meaning of unlawful messages and, as coded

language is also used in the political sphere, discourse analysts may do the same to unveil bias in the coded language of political figures and supporters. Finally, yet more applications involve e.g. political scientists and anthropologists, who can combine insights from the proposed analysis of semantic change with insights from psychology and cognitive science in order to better understand cultural transitions and the dynamic processes that shape them.

References

Hirotogu Akaike. 1998. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

John Langshaw Austin. 1975. *How to Do Things with Words*. Oxford University Press.

R Harald Baayen, Fabian Tomaschek, Susanne Gahl, and Michael Ramscar. 2017. The Ecclesiastes Principle in Language Change. *The changing English language: Psycholinguistic perspectives*, pages 21–48.

Robert Bamler and Stephan Mandt. 2017. Dynamic Word Embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR.org.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't Count, Predict! A Systematic Comparison of Context-counting vs. Context-predicting Semantic Vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Lisa Beinborn and Rochelle Choenni. 2019. Semantic Drift in Multilingual Representations. *arXiv preprint arXiv:1904.10820*.

Chris Biemann. 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*, pages 73–80. Association for Computational Linguistics.

Christopher M Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Andreas Blank and Peter Koch. 1999. Introduction: Historical Semantics and Cognition. *Historical Semantics and Cognition*, pages 1–16.

David M Blei and John D Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120. ACM.

Leonard Bloomfield. 1933. *Language*. New York: Allen & Unwin.

Gemma Boleda. 2019. Distributional Semantics and Linguistic Theory. *arXiv preprint arXiv:1905.01896*.

Gemma Boleda and Katrin Erk. 2015. Distributional Semantic Features as Semantic Primitivesor Not. In *2015 AAAI Spring Symposium Series*.

Michel Bréal. 1899. Essai de Sémantique (2e éd.). *Paris, Librairie Hachette et Cie*.

Claudia Marlea Brugman. 1988. *The Story of Over: Polysemy, Semantics, and the Structure of the Lexicon*. Garland, New York.

Joan Bybee. 2015. *Language Change*. Cambridge University Press.

Lyle Campbell. 2013. *Historical Linguistics*. Edinburgh University Press.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A Unified Model for Word Sense Representation and Disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. *arXiv preprint arXiv:1906.04341*.

Ronan Collobert and Jason Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel Word-sense Identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635.

Paul Cook and Suzanne Stevenson. 2010. Automatically Identifying the Source words of Lexical Blends in English. *Computational Linguistics*, 36(1):129–149.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised Sequence Learning. In *Advances in Neural Information Processing Systems*, pages 3079–3087.

Mark Davies. 2010. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and linguistic computing*, 25(4):447–464.

Mark Davies. 2012. Expanding Horizons in Historical Linguistics with the 400-million Word Corpus of Historical American English. *Corpora*, 7(2):121–157.

Marco Del Tredici and Raquel Fernández. 2018. The Road to Success: Assessing the Fate of Linguistic Innovations in Online Communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-Term Meaning Shift: A Distributional Exploration. In *Proceedings of NAACL-HLT 2019 (Annual Conference of the North American Chapter of the Association for Computational Linguistics)*.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A Bottom Up Approach to Category Mapping and Meaning Change. In *NetWordS*, pages 66–70.

Charles Elkan. 2003. Using the Triangle Inequality to Acceleratek-Means. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 147–153.

Katrin Erk and Sebastian Padó. 2008. A Structured Vector Space Model for Word Meaning in Context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based Model s for Word Meaning in Context. In *Proceedings of the acl 2010 conference short papers*, pages 92–97.

Stefan Evert. 2008. Corpora and Collocations. *Corpus Linguistics. An International Handbook*, 2:1212–1248.

Lea Frermann and Mirella Lapata. 2016. A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Charles Carpenter Fries. 1963. *Linguistics and Reading*, volume 2. Holt Rinehart and Winston.

Dirk Geeraerts et al. 1997. *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Oxford University Press.

Adele E Goldberg, Devin M Casenhiser, and Nitya Sethuraman. 2004. Learning Argument Structure Generalizations. *Cognitive Linguistics*, 15(3):289–316.

Stefan Th Gries and Anatol Stefanowitsch. 2004. Extending Collostructional Analysis: A Corpus-based Perspective on 'Alternations'. *International Journal of Corpus Linguistics*, 9(1):97–129.

SA Grondelaers, D Geeraerts, D Speelman, and H Cuyckens. 2007. Lexical Variation and Change. *Geeraerts, D.; Cuyckens, H.(ed.), The Oxford Handbook of Cognitive Linguistics*, pages 988–1011.

Kristina Gulordava and 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*, pages 67–71.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised Domain Adaptation of Contextualized Embeddings: A Case Study in Early Modern English. *arXiv preprint arXiv:1904.02817*.

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

Ruqaiya Hasan. 2009. *Semantic Variation: Meaning in Society and in Sociolinguistics*, volume 2. Equinox London.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hans Henrich Hock and Brian D Joseph. 2009. *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*, volume 218. Walter de Gruyter.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Information Technology*, 105:116.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and 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*, pages 61–65.

Walter Kintsch. 1988. The Role of Knowledge in Discourse Comprehension: A Construction-Integration Model. *Psychological Review*, 95(2):163.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic Word Embeddings and Semantic Shifts: a Survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 259–270.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word Sense Induction for Novel Sense Detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450*.

Omer Levy and Yoav Goldberg. 2014. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Jiwei Li and Dan Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732.

Jianhua Lin. 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical Word Embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Stuart Lloyd. 1982. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. OUP Oxford.

James MacQueen et al. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics.

Geoffrey McLachlan and Thriyambakam Krishnan. 2007. *The EM Algorithm and Extensions*, volume 382. John Wiley & Sons.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

James Milroy. 1992. *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. B. Blackwell, Oxford.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An Automatic Approach to Identify Word Sense Changes in Text Media Across Timescales. *Natural Language Engineering*, 21(5):773–798.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. Thats Sick Dude! Automatic Identification of Word Sense Change Across Different Timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029.

Franco Moretti. 2013. *Distant Reading*. Verso Books.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069.

Carita Paradis. 2011. Metonymization: A Key Mechanism in Semantic Change. *Defining Metonymy in Cognitive Linguistics: Towards a Consensus View*, pages 61–98.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Eethods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *arXiv preprint arXiv:1811.01088*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Michael Ramscar and Harald Baayen. 2013. Production, Comprehension, and synthesis: A Communicative Perspective on Language. *Frontiers in psychology*, 4:233.

Miguel A Ré and Rajeev K Azad. 2014. Generalization of Entropy Based Divergence Measures for Symbolic Sequence Analysis. *PloS one*, 9(4):e93532.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-Prototype Vector-Space Models of Word Meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics.

Alex Rosenfeld and Katrin Erk. 2018. Deep Neural Models of Semantic Shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484.

Maja Rudolph and David Blei. 2018. Dynamic Embeddings for Language Evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee.

Pavel Rychlỳ and Adam Kilgarriff. 2007. An Efficient Algorithm for Building a Distributional Thesaurus (and other Sketch Engine Developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 41–44. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. In *LCD Catalogue: LDC2008T19. DVD*. Philadelphia: Linguistic Data Consortium.

Lukas Schmelzeisen and Steffen Staab. 2019. Learning Taxonomies of Concepts and not Words using Contextualized Word Representations: A Position Paper. *arXiv preprint arXiv:1902.02169*.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123.

Gideon Schwarz et al. 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

John R Searle. 1975. A Taxonomy of Illocutionary Acts.

John R Searle. 1985. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics.

Anatol Stefanowitsch and Stefan Th Gries. 2003. Collostructions: Investigating the Interaction of Words and Constructions. *International Journal of Corpus Linguistics*, 8(2):209–243.

Gustaf Stern. 1931. Meaning and Change of Meaning with Special Reference to the English Language.

Terrence Szymanski. 2017. Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change Detection. *Computational Linguistics*, 1(1).

Wayne A Taylor. 2000. Change-point Analysis: A Powerful New Tool for Detecting Changes.

Wilson L Taylor. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Bulletin*, 30(4):415–433.

Elizabeth Closs Traugott. 2017. Semantic Change.

Elizabeth Closs Traugott and Richard B Dasher. 2001. *Regularity in Semantic Change*, volume 97. Cambridge University Press.

Elizabeth Closs Traugott and Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*, volume 6. Oxford University Press.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word Representations: a Simple and General Method for Semi-supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Peter D Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*, pages 35–40. ACM.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Zhaohui Wu and C Lee Giles. 2015. Sense-Aware Semantic Analysis: A Multi-Prototype Word Representation Model Using Wikipedia. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Yang Xu and Charles Kemp. 2015. A Computational Evaluation of Two Laws of Semantic Change. In *CogSci*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 19–27.

George Kingsley Zipf. 1949. Human behavior and the Principle of Least Effort.

Appendices

Appendix A

BERT Preprocessing

The pre-processing procedure to generate training examples for BERT involves the following steps.

1. Sentence segmentation: each document is split into the sentences that compose it. In the case of Reddit 2013, we consider each user post as a sentence. For non-conversational datasets, a sentence is usually simply marked by an end-of-sentence period; in fact, any segmentation algorithm can be used for this step.

2. Tokenisation: each sentence is split into tokens using the specific tokeniser that comes with the pre-trained BERT model: this involves punctuation splitting as well as the segmentation of word tokens into Wordpiece tokens

3. Sequence truncation: we set the maximum sequence length to 256 (a trade-off between modelling power and computational overhead) and truncate sequences at random both from the front and from the back, at random. That is, given a sequence of length $256 + m$, we remove the first or last token m times and with equal probability.

4. Sequence pair generation: as explained in Section 4.1, BERT's LM training involves a next sentence prediction task. So pairs are formed, with equal probability, either (i) by joining two actually contiguous segments or (ii) by sampling a sentence from a random document. The segment identifiers, for which segment embeddings are looked up, are 0 for the `[CLS]` token, the tokens forming the first sequence and the first `[SEP]`; they are 1 for the tokens of the second sequence and for the final `[SEP]` symbol.

5. Word masking: we mask Wordpiece tokens according to the default BERT probabilities (Section 4.1; Devlin et al., 2019).

6. Epoch-data generation: as the previous steps involve a significant random component, multiple epochs of pre-processed data can be generated to avoid training BERT on the same random split for all epochs.

Appendix B

Measures of inter-cluster distance

To establish links between usage type partitions obtained for two adjacent periods, the following inter-cluster distance measures can be used together with a distance metric d (e.g. Euclidean or cosine):

- *centroid distance*, the distance between the centroids of two clusters:

$$D_{CEN}\left(C_i, C_j\right) = d(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

- *single link distance*, the distance between the closest points of two clusters:

$$D_{SL}\left(C_i, C_j\right) = \min_{\boldsymbol{x}_i, \boldsymbol{x}_j} \{d(\boldsymbol{x}_i - \boldsymbol{x}_j) | \boldsymbol{x}_i \in C_i, \boldsymbol{x}_j \in C_j\}$$

- *complete link distance*, the distance between the furthest points of two clusters:

$$D_{CL}\left(C_i, C_j\right) = \max_{\boldsymbol{x}_i, \boldsymbol{x}_j} \{d(\boldsymbol{x}_i - \boldsymbol{x}_j) | \boldsymbol{x}_i \in C_i, \boldsymbol{x}_j \in C_j\}$$

- *average link distance*, the average pairwise distance between all possible combinations of points in the two clusters:

$$D_{AVG}\left(C_i, C_j\right) = \frac{1}{|C_i| \cdot |C_j|} \sum_{\substack{\boldsymbol{x}_i \in C_i, \\ \boldsymbol{x}_j \in C_j}} d(\boldsymbol{x}_i - \boldsymbol{x}_j)$$

- *Ward's distance*, the difference between the total intra-cluster sum of squares for the two clusters separately, and the intra-cluster sum of squares that results from merging the two clusters into a single cluster $C_{ij} = C_i \cup C_j$:

$$D_{WARD}\left(C_i, C_j\right) = \sum_{\boldsymbol{x}_i \in C_i} (\boldsymbol{x}_i - \boldsymbol{\mu}_i)^2 \; + \sum_{\boldsymbol{x}_j \in C_j} (\boldsymbol{x}_j - \boldsymbol{\mu}_j)^2 \; - \sum_{\boldsymbol{x}_{ij} \in C_{ij}} (\boldsymbol{x}_{ij} - \boldsymbol{\mu}_{ij})^2$$

Appendix C

Approximate change detection

As the thesis does not describe a way of tracking the entire vocabulary for semantic change, we propose here a simple method to do so approximately, which has not yet been thoroughly tested. For each word type occurring in the corpus, we keep track of its average contextualised representation and of its dimension-wise variance.

To make this process computationally feasible, we compute *mean-shift usage vectors* and *usage-variance vectors* incrementally. The mean-shift vector for word w after observing its n-th occurrence is given by:

$$\boldsymbol{\mu}_w^n = \mathrm{abs}\left(\boldsymbol{\mu}_w^n - \boldsymbol{\mu}_w^{n-1}\right)$$

$$\boldsymbol{m}_w^n = \boldsymbol{m}_w^{n-1} + \frac{\boldsymbol{x}_w^n - \boldsymbol{m}_w^{n-1}}{n}$$

where $\mathrm{abs}(\cdot)$ outputs a vector of element-wise absolute values. The usage-variance vector for word w after observing its n-th occurrence is given by:

$$\boldsymbol{\sigma}_w^n = \sqrt{\frac{\boldsymbol{s}_w^n}{n}}$$

$$\boldsymbol{s}_w^n = \boldsymbol{s}_w^{n-1} + \left(\boldsymbol{x}_w^n - \boldsymbol{\mu}_w^{n-1}\right)\left(\boldsymbol{x}_w^n - \boldsymbol{\mu}_w^n\right)$$

Meaning generalisation and specialisation should correspond to spikes and drops in variance, whereas reference shifts should correspond to changes in the mean usage vector.